EE566 ADAPTATION AND LEARNING **Instructor: Ali H. Sayed**
Homework Solutions #2 **Due: March 24, 2025**

---

1) (**Chapter 15**) Consider a first-order differentiable convex risk function $P(w)$ with bounded gradients, and introduce the convex optimization problem

$$w^\star = \underset{w\in\mathbb{R}^M}{\text{argmin}}\ P(w) \quad \text{subject to } w \in \mathcal{C}$$

where $\mathcal{C}$ is a convex set. The bounded gradient assumption amounts to

$$\|\nabla_w P(w)\|^2 \leq B, \quad \forall w \in \mathcal{C}$$

for some constant $B \geq 0$. Introduce the projection gradient algorithm

$$\begin{aligned} z_n &= w_{n-1} - \mu\nabla_{w^\mathsf{T}} P(w_{n-1}) \\ w_n &= \mathcal{P}_C[z_n] \end{aligned}$$

where $\mathcal{P}_C$ refers to the projection operator onto $\mathcal{C}$.

(a) Establish that

$$\frac{1}{N+1}\left(\sum_{m=0}^{N}\Big(P(w_{m-1}) - P(w^\star)\Big)\right) \leq \frac{\|\widetilde{w}_{-1}\|^2 + \mu^2 B^2(N+1)}{2(N+1)\mu}$$

(b) Assume we run the algorithm up to iteration N and return

$$w_N^{\text{best}} = \underset{0\leq n\leq N}{\text{argmin}}\ P(w_{n-1})$$

That is, we return the vector with the smallest risk value. Show that if $N = O(1/\epsilon^2)$ then

$$P(w_N^{\text{best}}) \leq P(w^\star) + \epsilon$$

(c) Assume we return instead the average weight

$$w_N^{\text{av}} = \frac{1}{N+1}\sum_{n=0}^{N} w_{n-1}$$

Show again that if $N = O(1/\epsilon^2)$ and $\mu = O(1/\sqrt{N})$ then

$$P(w_N^{\text{av}}) \leq P(w^\star) + \epsilon$$

Solution:

(a) Note that

$$\begin{aligned} w_n &= \mathcal{P}_c\Big[w_{n-1} - \mu\nabla_{w^\mathsf{T}} P(w_{n-1})\Big] \\ w^\star &= \mathcal{P}_C[w^\star] \end{aligned}$$

Using the nonexpansive property (9.70) we get

$$\begin{aligned} \|w^\star - w_n\|^2 &\leq \Big\|\mathcal{P}_C[w^\star] - \mathcal{P}_c\Big[w_{n-1} - \mu\nabla_{w^\mathsf{T}} P(w_{n-1})\Big]\Big\|^2 \\ &\leq \|w^\star - w_{n-1} + \mu\nabla_{w^\mathsf{T}} P(w_{n-1})\|^2 \end{aligned}$$

so that

$$\begin{aligned}\|\widetilde{w}_n\|^2 &\leq \|\widetilde{w}_{n-1} + \mu\nabla_{w^\intercal}P(w_{n-1})\|^2 \\ &\leq \|\widetilde{w}_{n-1}\|^2 + \mu^2\|\nabla_{w^\intercal}P(w_{n-1})\|^2 + 2\mu\nabla_w P(w_{n-1})\widetilde{w}_{n-1} \\ &\leq \|\widetilde{w}_{n-1}\|^2 + \mu^2 B^2 + 2\mu\nabla_w P(w_{n-1})\widetilde{w}_{n-1}\end{aligned}$$

Next, from convexity we have

$$P(w^\star) \geq P(w_{n-1}) + \nabla_w P(w_{n-1})(w^\star - w_{n-1})$$

That is,

$$\nabla_w P(w_{n-1})(w^\star - w_{n-1}) \leq P(w^\star) - P(w_{n-1})$$

It follows that

$$\|\widetilde{w}_n\|^2 \leq \|\widetilde{w}_{n-1}\|^2 + \mu^2 B^2 - 2\mu\Big(P(w_{n-1}) - P(w^\star)\Big)$$

Iterating gives

$$0 \leq \|\widetilde{w}_N\|^2 \leq \|\widetilde{w}_{-1}\|^2 + \mu^2 B^2(N+1) - 2\mu\sum_{m=0}^{N}\left(P(w_{m-1}) - P(w^\star)\right)$$

which implies that

$$\frac{1}{N+1}\left(\sum_{m=0}^{N}\Big(P(w_{m-1}) - P(w^\star)\Big)\right) \leq \frac{\|\widetilde{w}_{-1}\|^2 + \mu^2 B^2(N+1)}{2(N+1)\mu}$$

(b) For

$$w_N^{\text{best}} = \underset{0\leq n\leq N}{\text{argmin}}\ P(w_{n-1})$$

we have

$$P(w_N^{\text{best}}) - P(w^\star) \leq \frac{1}{N+1}\left(\sum_{m=0}^{N}\Big(P(w_{m-1}) - P(w^\star)\Big)\right)$$

and, hence,

$$P(w_N^{\text{best}}) - P(w^\star) \leq \frac{\|\widetilde{w}_{-1}\|^2}{2(N+1)}\frac{1}{\mu} + \frac{\mu B^2}{2}$$

Differentiating the upper bound relative to μ gives $\mu = O(1/\sqrt{N})$ from which

$$P(w_N^{\text{best}}) - P(w^\star) \leq O\Big(\frac{1}{\sqrt{N}}\Big)$$

For the upper bound to be $O(\epsilon)$ we need $N = O(1/\epsilon^2)$.

(c) By convexity we have

$$P(w_N^{\text{av}}) = P\Big(\frac{1}{N+1}\sum_{n=0}^{N} w_{n-1}\Big) \leq \frac{1}{N+1}\sum_{n=0}^{N} P(w_{n-1})$$

so that

$$P(w_N^{\text{av}}) - P(w^\star) \leq \frac{\|\widetilde{w}_{-1}\|^2 + \mu^2 B^2(N+1)}{2(N+1)\mu}$$

That is,

$$P(w_N^{\rm av}) - P(w^\star) \leq \frac{\|\widetilde{w}_{-1}\|^2}{2(N+1)}\frac{1}{\mu} + \frac{\mu B^2}{2}$$

Differentiating the upper bound relative to μ gives $\mu = O(1/\sqrt{N})$ from which

$$P(w_N^{\rm av}) - P(w^\star) \leq O\Big(\frac{1}{\sqrt{N}}\Big)$$

For the upper bound to be $O(\epsilon)$ we need $N = O(1/\epsilon^2)$.

□

2) (**Chapter 17**) Consider the following variation of Polyak's momentum acceleration method described originally by (17.65), where the second step involves now a combination of $\bar{\boldsymbol{b}}_{n-1}$ and $\boldsymbol{b}_n$ using a smoothing factor β:

$$\begin{aligned}
\boldsymbol{b}_n &= \nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) \\
\bar{\boldsymbol{b}}_n &= \beta\bar{\boldsymbol{b}}_{n-1} + (1-\beta)\boldsymbol{b}_n \\
\boldsymbol{w}_n &= \boldsymbol{w}_{n-1} - \mu\bar{\boldsymbol{b}}_n
\end{aligned}$$

(a) Assume $\boldsymbol{b}_n$ and $\bar{\boldsymbol{b}}_{n-1}$ are unbiased estimators for the true gradient vectors at $\boldsymbol{w}_{n-1}$ and $\boldsymbol{w}_{n-2}$, respectively, i.e., $\mathbb{E}\,\boldsymbol{b}_n = \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})$ and $\mathbb{E}\,\bar{\boldsymbol{b}}_{n-1} = \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-2})$. Is the updated vector $\bar{\boldsymbol{b}}_n$ an unbiased estimator for $\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})$?

(b) Assume we modify the second step as follows:

$$\bar{\boldsymbol{b}}_n = \beta\Big[\bar{\boldsymbol{b}}_{n-1} - \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-2}) + \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\Big] + (1-\beta)\boldsymbol{b}_n$$

Will $\bar{\boldsymbol{b}}_n$ be now an unbiased estimator for $\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_n)$?

(c) Use a Taylor series expansion to explain that the correction in part (b) can be approximated by the following in terms of the Hessian matrix of the empirical risk evaluated at $\boldsymbol{w}_{n-1}$:

$$\bar{\boldsymbol{b}}_n = \beta\Big[\bar{\boldsymbol{b}}_{n-1} + \mathbb{E}\,\nabla_w^2 P(\boldsymbol{w}_{n-1})(\boldsymbol{w}_{n-1} - \boldsymbol{w}_{n-2})\Big] + (1-\beta)\boldsymbol{b}_n$$

(d) What would be a stochastic approximation for the construction in part (c)?

Solution: For more details, the reader may refer to the works by Cutkosky, H. and F. Orabona (2019), "Momentum-based variance reduction in nonconvex SGD," *Proc. Advances Neural Information Processing Systems*, pp. 1–10, and Tran, H. and A. Cutkosky (2022), "Better SGD using second-order momentum" *Proc. Advances Neural Information Processing Systems*, pp. 3530–3541.

(a) Note that

$$\begin{aligned}
\mathbb{E}\,\bar{\boldsymbol{b}}_n &= \beta\mathbb{E}\,\bar{\boldsymbol{b}}_{n-1} + (1-\beta)\mathbb{E}\,\boldsymbol{b}_n \\
&= \beta\mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-2}) + (1-\beta)\mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1}) \\
&\neq \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})
\end{aligned}$$

It follows that the variable $\bar{\boldsymbol{b}}_n$ is not an unbiased estimator for $\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})$.

(b) In this case we have

$$\begin{aligned}
&\mathbb{E}\,\bar{\boldsymbol{b}}_n \\
&= \beta\Big[\mathbb{E}\,\bar{\boldsymbol{b}}_{n-1} - \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-2}) + \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\Big] + (1-\beta)\mathbb{E}\,\boldsymbol{b}_n \\
&= \beta\Big[\mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-2}) - \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-2}) + \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\Big] + (1-\beta)\mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1}) \\
&= \mathbb{E}\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})
\end{aligned}$$

It follows that $\bar{\boldsymbol{b}}_n$ is now an unbiased estimator for $\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})$?

(c) Using the Taylor series expansion we have

$$\nabla_{w^\mathsf{T}} P(w_{n-2}) = \nabla_{w^\mathsf{T}} P(w_{n-1}) + \nabla_w^2 P(w_{n-1})(w_{n-2} - w_{n-1}) + O(\|w_{n-2} - w_{n-1}\|^2)$$

Ignoring the higher order term and substituting into part (b) gives

$$\bar{\boldsymbol{b}}_n = \beta\Big[\bar{\boldsymbol{b}}_{n-1} + \mathbb{E}\,\nabla_w^2 P(\boldsymbol{w}_{n-1})(\boldsymbol{w}_{n-1} - \boldsymbol{w}_{n-2})]\Big] + (1-\beta)\boldsymbol{b}_n$$

(d) A stochastic approximation replaces the Hessian matrix of $P(w)$ by the Hessian matrix of the loss function to get

$$\bar{\boldsymbol{b}}_n = \beta\Big[\bar{\boldsymbol{b}}_{n-1} + \nabla_w^2 Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n)(\boldsymbol{w}_{n-1} - \boldsymbol{w}_{n-2})]\Big] + (1-\beta)\boldsymbol{b}_n$$

□

3) (**Chapter 18**) Consider a collection $\mathcal{S} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ of N independent random vectors, each with mean $\mathbb{E}\boldsymbol{x}_n = \overline{x}_n$ and variance $\sigma_n^2 = \mathbb{E}\|\boldsymbol{x}_n - \mathbb{E}\boldsymbol{x}_n\|^2$. Introduce the sample mean

$$\overline{x} \triangleq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$$

We wish to estimate $\overline{x}$ by considering two constructions based on a mini-batch of size B. The first construction selects B at random *without* replacement from $\mathcal{S}$, while the second construction selects B samples *with* replacement. We do not assume uniform sampling. Instead, we let p_n denote the normalized *inclusion* probability of $\boldsymbol{x}_n$ in the mini-batch $\mathcal{B}$. We denote the samples selected with replacement by $\{\boldsymbol{x}_b^r\}$ and the samples selected without replacement by $\{\boldsymbol{x}_b^{wr}\}$. We construct the estimators as follows:

$$\widehat{\boldsymbol{x}}^{\mathrm{r}} \triangleq \frac{1}{B}\sum_{b=1}^{B} \frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}} \qquad (\textit{with} \text{ replacement})$$

$$\widehat{\boldsymbol{x}}^{\mathrm{wr}} \triangleq \frac{1}{B}\sum_{b=1}^{B} \frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{wr}} \qquad (\textit{without} \text{ replacement})$$

Show that both estimators are unbiased with variances

$$\mathbb{E}\|\widehat{\boldsymbol{x}}^{\mathrm{r}} - \overline{x}\|^2 = \frac{1}{B}\sum_{n=1}^{N} p_n \left(\frac{1}{N^2 p_n^2}\sigma_n^2 + \left\|\frac{1}{Np_n}\overline{x}_n - \overline{x}\right\|^2\right)$$

$$\mathbb{E}\|\widehat{\boldsymbol{x}}^{\mathrm{wr}} - \overline{x}\|^2 \leq \frac{1}{B}\sum_{n=1}^{N} p_n \left(\frac{1}{N^2 p_n^2}\sigma_n^2 + \left\|\frac{1}{Np_n}\overline{x}_n - \overline{x}\right\|^2\right)$$

Solution: The argument appears in the work by Rizk, E., S. Vlaski, and A. H. Sayed (2022), "Federated learning under importance sampling," *IEEE Trans. Signal Process.,* vol. 70, pp. 5381–5396. The derivation extends a result from Brewer, K. R. W. and M. Hanif (1983), *Sampling with Unequal Probabilities*, Springer-Verlag, NY. The presentation is adjusted for the current notation and context.

We begin with the *with-replacement* setting. The randomness of the samples introduces some intri-

cacies that need to be accounted for in the notation. For the mean, we have:

$$\begin{aligned}
\mathbb{E}\widehat{\boldsymbol{x}}^{\mathrm{r}} &= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left(\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}\right)\\
&= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left\{\mathbb{E}\left\{\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}\middle|\mathcal{S}\right\}\right\}\\
&= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left\{\sum_{n=1}^{N}p_n\frac{1}{Np_n}\boldsymbol{x}_n\right\} = \frac{1}{B}\sum_{b=1}^{B}\overline{x}\\
&= \overline{x}.
\end{aligned}$$

For the variance we find:

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{x}}^{\mathrm{r}}-\overline{x}\|^2\\
&= \mathbb{E}\left\|\frac{1}{B}\sum_{b=1}^{B}\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right\|^2\\
&= \mathbb{E}\left\|\frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right)\right\|^2,\\
&= \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\|\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right\|^2\\
&\quad+\frac{1}{B^2}\sum_{b_1\neq b_2}\mathbb{E}\left\{\left(\frac{1}{Np_{b_1}}\boldsymbol{x}_{b_1}-\overline{x}\right)\left(\frac{1}{Np_{b_2}}\boldsymbol{x}_{b_2}-\overline{x}\right)\right\}\\
&\overset{(a)}{=} \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\|\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right\|^2\\
&\quad+\frac{1}{B^2}\sum_{b_1\neq b_2}\mathbb{E}\left\{\frac{1}{Np_{b_1}}\boldsymbol{x}_{b_1}-\overline{x}\right\}\mathbb{E}\left\{\frac{1}{Np_{b_2}}\boldsymbol{x}_{b_2}-\overline{x}\right\},\\
&\overset{(b)}{=} \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\|\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right\|^2
\end{aligned}$$

where (a) is a result of the fact that the elements of $\mathcal{S}$ are independent and $\boldsymbol{x}_b^{\mathrm{r}}$ is sampled from $\mathcal{S}$ independently, and hence $\boldsymbol{x}_{b_1}$ and $\boldsymbol{x}_{b_2}$ are independent. Step (b) follows from:

$$\mathbb{E}\left(\frac{1}{Np_b}\boldsymbol{x}_b\right) = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\right) = \overline{x}.$$

Then,

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{x}}^{\mathrm{r}}-\overline{x}\|^2\\
&=\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\|\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right\|^2\\
&=\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\mathbb{E}\left\|\frac{1}{Np_b}\boldsymbol{x}_b^{\mathrm{r}}-\overline{x}\right\|^2\Big|\mathcal{S}\right\}\\
&=\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\sum_{n=1}^{N}p_n\left\|\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right\|^2\right\},\\
&=\frac{1}{B^2}\sum_{b=1}^{B}\sum_{n=1}^{N}p_n\mathbb{E}\left\|\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right\|^2\\
&=\frac{1}{B}\sum_{n=1}^{N}p_n\mathbb{E}\left\|\frac{1}{Np_n}\boldsymbol{x}_n-\frac{1}{Np_n}\overline{x}_n+\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\\
&=\frac{1}{B}\sum_{n=1}^{N}p_n\left(\mathbb{E}\left\|\frac{1}{Np_n}\boldsymbol{x}_n-\frac{1}{Np_n}\overline{x}_n\right\|^2+\left\|\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\right)\\
&=\frac{1}{B}\sum_{n=1}^{N}p_n\left(\frac{1}{N^2p_n^2}\sigma_n^2+\left\|\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\right).
\end{aligned}$$

We now proceed to study the *without replacement* setting. The fact that the $\boldsymbol{x}_b$ are sampled from $\mathcal{S}$ without replacement causes pairs $\boldsymbol{x}_{b_1},\boldsymbol{x}_{b_2}$ to no longer be independent. We denote the set of points sampled from $\mathcal{S}$ *without replacement* by $\mathcal{B}^{\mathrm{wr}}$ and introduce the activation function by:

$$\mathbb{I}_n\triangleq\begin{cases}1,\ \text{if }\boldsymbol{x}_n\in\mathcal{B}^{\mathrm{wr}}\\0,\ \text{if }\boldsymbol{x}_n\notin\mathcal{B}^{\mathrm{wr}}\end{cases}$$

Then, the estimator $\widehat{\boldsymbol{x}}^{\mathrm{wr}}$ can be written equivalently as:

$$\widehat{\boldsymbol{x}}^{\mathrm{wr}}=\frac{1}{B}\sum_{n=1}^{N}\mathbb{I}_n\frac{1}{Np_n}\boldsymbol{x}_n$$

For the mean, we have:

$$\begin{aligned}
\mathbb{E}\widehat{\boldsymbol{x}}^{\mathrm{wr}}&=\frac{1}{B}\sum_{n=1}^{N}\mathbb{E}\left\{\mathbb{I}_n\frac{1}{Np_n}\boldsymbol{x}_n\right\}=\frac{1}{B}\sum_{n=1}^{N}\mathbb{E}\mathbb{I}_n\times\mathbb{E}\frac{1}{Np_n}\boldsymbol{x}_n\\
&=\frac{1}{B}\sum_{n=1}^{N}Bp_n\times\frac{1}{Np_n}\overline{x}_n=\frac{1}{N}\sum_{n=1}^{N}\overline{x}_n=\overline{x}
\end{aligned}$$

For the variance, we have:

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{x}}^{\mathrm{wr}}-\overline{x}\right\|^2 &= \mathbb{E}\left\|\frac{1}{B}\sum_{n=1}^{N}\mathbb{I}_n\left(\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right)\right\|^2\\ &= \frac{1}{B^2}\sum_{n=1}^{N}\mathbb{E}\left\|\mathbb{I}_n\left(\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right)\right\|^2\\ &\quad+\frac{1}{B^2}\sum_{n_1\neq n_2}\mathbb{E}\Bigg\{\mathbb{I}_{n_1}\left(\frac{1}{Np_{n_1}}\boldsymbol{x}_{n_1}-\overline{x}\right)\mathbb{I}_{n_2}\\ &\qquad\qquad\times\left(\frac{1}{Np_{n_2}}\boldsymbol{x}_{n_2}-\overline{x}\right)\Bigg\}\end{aligned}$$

We begin with:

$$\begin{aligned}&\mathbb{E}\left\|\mathbb{I}_n\left(\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right)\right\|^2\\ &=\mathbb{E}\left\{\left\|\mathbb{I}_n\left(\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right)\right\|^2\Big|\mathbb{I}_n=1\right\}\times\mathbb{P}\left(\mathbb{I}_n=1\right)\\ &\quad+\mathbb{E}\left\{\left\|\mathbb{I}_n\left(\frac{1}{Np_n}\boldsymbol{x}_n-\overline{x}\right)\right\|^2\Big|\mathbb{I}_n=0\right\}\times\mathbb{P}\left(\mathbb{I}_n=0\right)\\ &=Bp_n\left(\mathbb{E}\left\|\frac{1}{Np_n}\boldsymbol{x}_n-\frac{1}{Np_n}\overline{x}_n+\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\right)\\ &=Bp_n\left(\frac{1}{N^2p_n^2}\mathbb{E}\|\boldsymbol{x}_n-\overline{x}_n\|^2+\left\|\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\right)\\ &=Bp_n\left(\frac{1}{N^2p_n^2}\sigma_n^2+\left\|\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\right)\end{aligned}$$

For the cross-term we have:

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{I}_{n_1}\left(\frac{1}{Np_{n_1}}\boldsymbol{x}_{n_1}-\overline{x}\right)\mathbb{I}_{n_2}\left(\frac{1}{Np_{n_2}}\boldsymbol{x}_{n_2}-\overline{x}\right)\right\}\\ &=\mathbb{E}\Bigg\{\left(\frac{1}{Np_{n_1}}\boldsymbol{x}_{n_1}-\overline{x}\right)\left(\frac{1}{Np_{n_2}}\boldsymbol{x}_{n_2}-\overline{x}\right)\Big|\mathbb{I}_{n_1}=1,\mathbb{I}_{n_2}=1\Bigg\}\\ &\quad\times\mathbb{P}\left(\mathbb{I}_{n_1}=1,\mathbb{I}_{n_2}=1\right)\\ &=\mathbb{P}\left(\mathbb{I}_{n_2}=1,\mathbb{I}_{n_1}=1\right)\left(\frac{1}{Np_{n_1}}\mathbb{E}\boldsymbol{x}_{n_1}-\overline{x}\right)\left(\frac{1}{Np_{n_2}}\mathbb{E}\boldsymbol{x}_{n_2}-\overline{x}\right)\\ &=\mathbb{P}\left(\mathbb{I}_{n_2}=1,\mathbb{I}_{n_1}=1\right)\left(\frac{1}{Np_{n_1}}\overline{x}_{n_1}-\overline{x}\right)\left(\frac{1}{Np_{n_2}}\overline{x}_{n_2}-\overline{x}\right)\end{aligned}$$

We then get the desired result:

$$\begin{aligned}
&\mathbb{E}\,\|\widehat{\boldsymbol{x}}^{\mathrm{wr}}-\overline{x}\|^2\\
&=\frac{1}{B}\sum_{n=1}^{N}p_n\left(\frac{1}{N^2p_n^2}\sigma_n^2+\left\|\frac{1}{Np_n}\overline{x}_n-\overline{x}\right\|^2\right)\\
&\quad+\frac{1}{B^2}\sum_{n_1\neq n_2}\mathbb{P}\left\{\mathbb{I}_{n_2}=1,\mathbb{I}_{n_1}=1\right\}\left(\frac{1}{Np_{n_1}}\overline{x}_{n_1}-\overline{x}\right)\\
&\quad\times\left(\frac{1}{Np_{n_2}}\overline{x}_{n_2}-\overline{x}\right).
\end{aligned}$$

□

4) (**Chapter 19**) Consider an empirical risk minimization problem of the following form where $P(w)$ is $\nu-$strongly convex and has $\delta-$Lipschitz gradients:

$$w^\star\;\triangleq\;\underset{w\in\mathbb{R}^M}{\operatorname{argmin}}\left\{P(w)\;\triangleq\;\frac{1}{N}\sum_{m=0}^{N-1}Q(w;\gamma(m),h_m)\right\}$$

Consider the average regret from the body of the chapter defined as

$$\mathbb{E}\,\boldsymbol{\mathcal{R}}(N)\;\triangleq\;\frac{1}{N}\sum_{n=0}^{N-1}\Big(\mathbb{E}\,P(\boldsymbol{w}_{n-1})-P(w^\star)\Big)$$

The empirical risk is minimized by using the classical stochastic gradient algorithm with step-size μ and where the gradient noise satisfies conditions (19.13a)–(19.13b).

(a) Show that an approximate upper bound for $\mathbb{E}\,\boldsymbol{\mathcal{R}}(N)$ is given by

$$\mathbb{E}\,\boldsymbol{\mathcal{R}}(N)\leq\frac{\delta}{4N\mu\nu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2\delta}{4\nu}$$

(b) Minimize the upper bound with respect to μ and conclude that $\mathbb{E}\,\boldsymbol{\mathcal{R}}(N)\leq O(1/\sqrt{N})$.

Solution:

(a) We know that

$$\begin{aligned}
&0\leq\mathbb{E}\,P(\boldsymbol{w}_{n-1})-P(w^\star)\leq\frac{\delta}{2}\|\widetilde{\boldsymbol{w}}_{n-1}\|^2\\
&\|\mathbb{E}\,\widetilde{\boldsymbol{w}}_{n-1}\|^2\leq\lambda^n\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu^2\sigma_g^2}{1-\lambda}\approx\lambda^n\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2}{2\nu}
\end{aligned}$$

where we approximated $1-\lambda\approx2\mu\nu$. It follows that

$$\mathbb{E}\,P(\boldsymbol{w}_{n-1})-P(w^\star)\leq\frac{\delta}{2}\lambda^n\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2\delta}{4\nu}$$

and, hence,

$$\begin{aligned}\mathbb{E}\,\boldsymbol{\mathcal{R}}(N) &\leq \frac{1}{N}\sum_{n=0}^{N-1}\Big(\frac{\delta}{2}\lambda^n\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2\delta}{4\nu}\Big)\\ &\leq \frac{\delta}{2N}\frac{1}{1-\lambda}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2\delta}{4\nu}\\ &\approx \frac{\delta}{2N}\frac{1}{2\mu\nu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2\delta}{4\nu}\\ &= \frac{\delta}{4N\mu\nu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\mu\sigma_g^2\delta}{4\nu}\end{aligned}$$

(b) Differentiating the upper bound relative to μ and setting the derivative to 0 gives:

$$-\frac{\delta}{4N\nu}\frac{1}{\mu^2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2+\frac{\sigma_g^2\delta}{4\nu}=0$$

for which

$$\mu^o=\frac{\Big(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2\Big)^{1/2}}{\sigma_g\sqrt{N}}$$

Substituting into the upper bound for $\mathbb{E}\,\boldsymbol{\mathcal{R}}(N)$ gives

$$\mathbb{E}\,\boldsymbol{\mathcal{R}}(N)\leq\frac{\delta\sigma_g\Big(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2\Big)^{1/2}}{2\nu}\times\frac{1}{\sqrt{N}}\;=\;O(1/\sqrt{N})$$

□

5) (**Chapter 27**) Consider the noisy observation $\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{v}$ where all variables are $M-$dimensional and $\boldsymbol{v}$ is Gaussian noise with zero mean and covariance matrix R_v, i.e., $\boldsymbol{v}\sim\mathcal{N}(0,R_v)$. The noise and $\boldsymbol{x}$ are independent of each other and R_v is positive definite.

(a) Let $f_{\boldsymbol{x}}(x)$ denote the pdf of $\boldsymbol{x}$. Argue that the pdf of $\boldsymbol{y}$ is given by

$$f_{\boldsymbol{y}}(y)\;\triangleq\;\int_{x\in\mathcal{X}}\frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}f_{\boldsymbol{x}}(x)dx$$

Argue further that the conditional pdf of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)=\frac{f_{\boldsymbol{x}}(x)}{f_{\boldsymbol{y}}(y)}\times\frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}$$

(b) Verify that the gradient of $f_{\boldsymbol{y}}(y)$ relative to y satisfies

$$\frac{1}{f_{\boldsymbol{y}}(y)}\nabla_{y^{\mathsf{T}}}f_{\boldsymbol{y}}(y)=-R_v^{-1}(y-\widehat{x})$$

where $\widehat{x}=\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y)$.

(c) Conclude that the least mean-squares estimator satisfies the following relation in terms of the pdf of the observation:

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})=\boldsymbol{y}+R_v\nabla_{y^{\mathsf{T}}}\ln f_{\boldsymbol{y}}(y)$$

Solution:

(a) From Bayes rule

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y)=f_{\boldsymbol{x}}(x)f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)=f_{\boldsymbol{y}}(y)f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$$

Moreover, we have that

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) = f_{\boldsymbol{v}}(y-x) = \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}$$

Therefore,

$$\begin{aligned} f_{\boldsymbol{x},\boldsymbol{y}}(x,y) &= f_{\boldsymbol{x}}(x)f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \\ &= f_{\boldsymbol{x}}(x) \times \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\} \end{aligned}$$

from which we conclude by marginalizing over x that

$$f_{\boldsymbol{y}}(y) = \int_{x\in\mathcal{X}} \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}f_{\boldsymbol{x}}(x)dx$$

Likewise, from Bayes rule again

$$\begin{aligned} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) &= \frac{f_{\boldsymbol{x}}(x)f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)}{f_{\boldsymbol{y}}(y)} \\ &= \frac{f_{\boldsymbol{x}}(x)}{f_{\boldsymbol{y}}(y)} \times \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\} \end{aligned}$$

(b) Differentiating $f_{\boldsymbol{y}}(y)$ relative to y and switching differentiation with integration gives

$$\begin{aligned} &\nabla_{y^{\mathsf{T}}} f_{\boldsymbol{y}}(y) \\ &= \nabla_{y^{\mathsf{T}}}\left\{\int_{x\in\mathcal{X}} \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}f_{\boldsymbol{x}}(x)dx\right\} \\ &= \int_{x\in\mathcal{X}} \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\nabla_{y^{\mathsf{T}}}\left\{\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}\right\}f_{\boldsymbol{x}}(x)dx \\ &= -\frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\int_{x\in\mathcal{X}} R_v^{-1}(y-x)\exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}f_{\boldsymbol{x}}(x)dx \\ &= -R_v^{-1} \times \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\int_{x\in\mathcal{X}} y \times \exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}f_{\boldsymbol{x}}(x)dx + \\ &\quad R_v^{-1} \times \frac{1}{\sqrt{(2\pi)^M}}\frac{1}{\sqrt{\det R_v}}\int_{x\in\mathcal{X}} x \times \exp\Big\{-\frac{1}{2}(y-x)^{\mathsf{T}}R_v^{-1}(y-x)\Big\}f_{\boldsymbol{x}}(x)dx \\ &= -R_v^{-1} \times \int_{x\in\mathcal{X}} y \times f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)f_{\boldsymbol{x}}(x)dx + R_v^{-1} \times f_{\boldsymbol{y}}(y)\int_{x\in\mathcal{X}} x \times f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \\ &= -R_v^{-1} \times \int_{x\in\mathcal{X}} y \times f_{\boldsymbol{x},\boldsymbol{y}}(x,y)dx + R_v^{-1} \times f_{\boldsymbol{y}}(y) \times \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) \\ &= -R_v^{-1} \times f_{\boldsymbol{y}}(y) \times \int_{x\in\mathcal{X}} y \times f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx + R_v^{-1} \times f_{\boldsymbol{y}}(y)\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) \\ &= -R_v^{-1} \times f_{\boldsymbol{y}}(y) \times y \times \underbrace{\int_{x\in\mathcal{X}} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx}_{=1} + R_v^{-1} \times f_{\boldsymbol{y}}(y)\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) \\ &= R_v^{-1} \times f_{\boldsymbol{y}}(y) \times \Big(\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) - y\Big) \end{aligned}$$

Therefore,

$$\frac{1}{f_{\boldsymbol{y}}(y)}\nabla_{y^{\mathsf{T}}} f_{\boldsymbol{y}}(y) = R_v^{-1} \times f_{\boldsymbol{y}}(y) \times \Big(\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) - y\Big)$$

where $\widehat{x} = \mathbb{E}(\boldsymbol{x}|\boldsymbol{y} = y)$.

(c) Rearranging terms and noting that

$$\nabla_{y^\intercal} \ln f_{\boldsymbol{y}}(y) = \frac{1}{f_{\boldsymbol{y}}(y)} \nabla_{y^\intercal} f_{\boldsymbol{y}}(y)$$

we conclude that

$$\mathbb{E}(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{y} + R_v \nabla_{y^\intercal} \ln f_{\boldsymbol{y}}(y)$$

□