Adaptation, Learning, and Optimization over Networks

Ali H. Sayed
University of California at Los Angeles

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. H. Sayed. *Adaptation, Learning, and Optimization over Networks*. Foundations and Trends® in Machine Learning, vol. 7, no. 4-5, pp. 311–801, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-850-8

Problems

This version was posted online on August 13, 2014.

This list of problems is provided by the author, A. H. Sayed (sayed@ee.ucla.edu), for educational purposes only in order to complement the text listed below, which is used as a reference in his graduate-level course *Inference over Networks* at UCLA. A solutions manual is available to instructors upon request:

A. H. Sayed, *Adaptation, Learning, and Optimization over Networks*, Foundations and Trends in Machine Learning, volume 7, issue 4-5, NOW Publishers, Boston-Delft, 518pp, 2014.

ISBN 978-1-60198-850-8, DOI 10.1561/2200000051.

This list is an evolving collection of assignment problems and is expected to undergo regular updates. Please excuse imperfections. Please visit this page on the author's website to check for updates.

APPENDICES A and B

P 1. Let $g(x,z) = x^{\mathsf{T}}Cz$, where $x, z \in \mathbb{R}^M$ and C is a matrix. Verify that

$$\nabla_z\, g(x,z) = x^{\mathsf{T}}C, \quad \nabla_x\, g(x,z) = z^{\mathsf{T}}C^{\mathsf{T}}$$

P 2. Let $g(z) = x^{\mathsf{T}}Cw$, where $x, w \in \mathbb{R}^M$ and both are functions of a vector $z \in \mathbb{R}^P$, i.e., $x = x(z)$ and $w = w(z)$, and C is a matrix that is independent of z. Establish the chain rule

$$\nabla_z\, g(z) = x^{\mathsf{T}}C\,(\nabla_z\, w(z)) \;+\; w^{\mathsf{T}}C^{\mathsf{T}}\,(\nabla_z\, x(z))$$

where, since $x(z)$ is a vector-valued function of a vector argument, its gradient (or Jacobian) relative to z is now defined as the $M \times P$ matrix function whose entries are given by:

$$[\nabla_z\, x(z)]_{m,p} \;\stackrel{\Delta}{=}\; \frac{\partial x_m}{\partial z_n}, \quad m = 1, 2, \ldots, M, \;\; p = 1, 2, \ldots, P$$

That is, the $(m,p)-$th entry is equal to the partial derivative of the $m-$th entry of the vector $x(z)$ relative to the $p-$th entry of z. Similarly for $\nabla_z\, w(z)$.

P 3. Let $g(z)$ be a real-valued differentiable function with $z \in \mathbb{R}^M$. Assume the entries of z are functions of a scalar parameter t,

$$z = \text{col}\{z_1(t),\, z_2(t), \ldots, z_M(t)\}$$

Introduce the column vector

$$\frac{dz}{dt} \;\stackrel{\Delta}{=}\; \text{col}\left\{\frac{dz_1(t)}{dt}, \frac{dz_2(t)}{dt}, \ldots, \frac{dz_M(t)}{dt}\right\}$$

Show that

$$\frac{dg}{dt} \;=\; [\nabla_z\, g(z)]\,\frac{dz}{dt}$$

P 4. Let $g(z)$ be a real-valued function with $z \in \mathbb{R}^M$. Let $f(t)$ be a real-valued function with $t \in \mathbb{R}$. Both functions are differentiable in their arguments. Show that

$$\nabla_z\, f(g(z)) \;=\; \left(\left.\frac{df(t)}{dt}\right|_{t=g(z)}\right) \nabla_z\, g(z)$$

P 5. Let $g(z)$ be a real-valued twice-differentiable function with $z \in \mathbb{R}^M$. Define $f(t) = g(z + t\Delta z)$ for $t \in [0,1]$. Show from first principles that (D.4) holds, namely,

$$\frac{df(t)}{dt} \;=\; [\nabla_z\, g(z + t\Delta z)]\,\Delta z$$

Show also that

$$\frac{d^2 f(t)}{dt^2} \;=\; (\Delta z)^{\mathsf{T}} \left[\nabla_z^2\, g(z+t\Delta z)\right] \Delta z$$

P 6. Compute the gradient vectors relative to z and the Hessian matrices of the following functions for $z \in \mathbb{C}^{M\times 1}$:

(a) $g(z) = z^{\mathsf{T}} z$.

(b) $g(z) = \|2\text{Re}(z)\|^2$.

(c) $g(z) = \|z\|^4$.

P 7. Consider $g(z) = \|z\|^4 + 2\|\text{Re}(z)\|^2$, where $z \in \mathbb{C}^M$.

(a) Determine the complex gradients $\nabla_z\, g(z)$, $\nabla_{z^*} g(z)$, and $\nabla_{z^{\mathsf{T}}} g(z)$.

(b) Determine $H(v)$, the real-form of the complex Hessian matrix.

(c) Determine $H_c(u)$, the complex-form of the complex Hessian matrix.

(d) Verify that the Hessian matrices of parts (b)-(c) satisfy relation (B.26).

(e) Is $g(z)$ convex? strictly convex? strongly-convex?

(f) Find the gradient vector and Hessian matrix when $z \in \mathbb{R}^M$.

P 8. Establish the validity of relation (B.26).

APPENDIX C

P 9. Show that definitions (C.3) and (C.4) are equivalent characterizations of convexity when $g(z)$ is differentiable.

P 10. Establish property (2) in Example C.2.

P 11. Let $g(z) \in \mathbb{R}$ and $z \in \mathbb{C}^M$. Show that $g(z) = \|z\|^4$ is strictly convex.

P 12. Show that the regularized hinge loss function (C.24) is strongly convex.

P 13. Establish (C.40) as an equivalent characterization for $\nu-$strong convexity for functions $g(z) \in \mathbb{R}$ of complex arguments $z \in \mathbb{C}^M$.

P 14. Establish property (C.43) for $\nu-$strongly convex functions $g(z) \in \mathbb{R}$ of complex arguments $z \in \mathbb{C}^M$.

P 15. Let $z \in \mathbb{C}^M$ and consider a full-rank matrix $A \in \mathbb{C}^{N\times M}$ with $N \geq M$. Examine the convexity, strict convexity, and strong-convexity of the function $g(z) = \|Az\|^\alpha$ for all values of α in the range $\alpha \in [1,\infty)$. How would your answers change if A were nonzero but rank-deficient?

APPENDICES D and E

P 16. Consider $g(z) = \|z\|^4 + \|2\text{Re}(z)\|^2$, where $z \in \mathbb{C}^M$. Let z^o denote a stationary point of $g(z)$ and introduce $\widetilde{z} = z^o - z$. Use the mean-value theorem (D.19) to express the difference $g(z) - g(z^o)$ in terms of $\widetilde{z}$. Write down expression (D.20) for this case as well. Evaluate the integral expressions whenever possible.

P 17. Let $g(z)$ be a real-valued twice-differentiable function with $z \in \mathbb{R}^M$. Use the result of Lemma D.1 to conclude that

$$g(z_o+\Delta z) = g(z_o)+[\nabla_z\, g(z_o)]\,\Delta z + (\Delta z)^{\mathsf{T}}\left(\int_0^1\int_0^1 t\nabla_z^2\, g(z_o + tr\Delta z)drdt\right)\Delta z$$

How would this result change if $z \in \mathbb{C}^M$?

P 18. Consider column vectors $h, z \in \mathbb{R}^M$ and $\rho > 0$. Introduce the logistic function:

$$g(z) \;=\; \frac{\rho}{2}\|z\|^2 \;+\; \ln\left(1 + e^{-h^{\mathsf{T}}z}\right)$$

(a) Show that $g(z)$ is strongly-convex. Let z^o denote its global minimizer.

(b) Show that, for any z, the Hessian matrix function of $g(z)$ satisfies a Lipschitz condition of the form

$$\left\|\nabla_z^2\, g(z) - \nabla_z^2\, g(z^o)\right\| \;\leq\; \kappa\,\|z^o - z\|$$

for some $\kappa \geq 0$. Determine an expression for κ in terms of h.

P 19. Problems 19–21 are motivated by useful properties from [190][Ch. 1]. Consider a convex function $g(z) \in \mathbb{R}$ with a gradient vector that satisfies the Lipschitz condition (E.21) with $z \in \mathbb{R}^M$. Let z^o denote the location of a global minimum for $g(z)$ and define $\widetilde{z} = z^o - z$. Show that:

$$(\nabla_z\, g(z))\,\widetilde{z} \;\leq\; -(1/\delta)\,\|\nabla_z\, g(z)\|^2$$

Provide one interpretation for this result.

P 20. Consider a $\nu-$strongly convex function $g(z) \in \mathbb{R}$ with $z \in \mathbb{R}^M$. Let z^o denote its unique global minimum. Show that

$$g(z) - g(z^o) \;\leq\; (1/2\nu)\,\|\nabla_z\, g(z)\|^2$$

How does this result relate to expression (E.17).

P 21. Let $g(z)$ be a real-valued twice-differentiable function with $z \in \mathbb{R}^M$. Refer to the mean-value theorem (D.8).

(a) Assume the gradient of $g(z)$ is $\delta-$Lipschitz continuous in the interval $z \in [z_o, z_o + \Delta z]$ so that

$$\|\nabla_z\, g(z_o + t\Delta z) - \nabla_z\, g(z_o)\| \leq \delta \cdot t \cdot \|\Delta z\|$$

for any $t \in [0,1]$. Show that

$$\|g(z_o + \Delta z) - g(z_o) - [\nabla_z\, g(z_o)]\, \Delta z\| \;\leq\; \frac{\delta}{2}\|\Delta z\|^2$$

(b) Assume instead that the Hessian matrix of $g(z)$ is $\delta-$Lipschitz continuous in the interval $z \in [z_o, z_o + \Delta z]$, i.e.,

$$\|\nabla_z^2\, g(z_o + t\Delta z) - \nabla_z^2\, g(z_o)\| \leq \delta \cdot t \cdot \|\Delta z\|$$

for any $t \in [0,1]$. Show that

$$\left\|g(z_o + \Delta z) - g(z_o) - [\nabla_z\, g(z_o)]\, \Delta z - \frac{1}{2}\, (\Delta z)^{\mathsf{T}} \left[\nabla_z^2\, g(z_o)\right]\, \Delta z\right\| \;\leq\; \frac{\delta}{6}\|\Delta z\|^3$$

APPENDICES F and G

P 22. Establish the validity of the third property in Table F.2.

P 23. Establish the singular value property (8) for Kronecker products from Table F.1.

P 24. Consider matrices A, B, C, and D of compatible dimensions. Show that

$$\text{Tr}(A^{\mathsf{T}} BCD^{\mathsf{T}}) \;=\; (\text{vec}(A))^{\mathsf{T}}\, (D \otimes B)\text{vec}(C)$$

P 25. Show that $\text{Tr}(A \otimes B) = \text{Tr}(A)\text{Tr}(B)$.

P 26. Verify that when B is Hermitian, it also holds that $\text{Tr}(AB) = [\text{vec}(B)]^*\, \text{vec}(A)$.

P 27. Show that, for any matrix norm, $|\text{Tr}(A)| \leq c \cdot \|A\|$ for some constant c.

P 28. Establish the validity of (F.5), namely, that the $2-$induced norm of a matrix is equal to its maximum singular value.

P 29. Let all eigenvalues of $A \in \mathbb{C}^{N\times N}$ have negative real parts. The matrix A is not necessarily Hermitian and, therefore, its eigenvalues can be complex. Find a condition on $\mu > 0$ to ensure that the matrix $I_N + \mu A$ is stable.

P 30. For every matrix A, any matrix norm, and any $\epsilon > 0$, show that it holds:

$$\|A^n\| \;\leq\; c(\rho(A) + \epsilon)^n$$

P 31. For any matrix norm, show that the spectral radius of a matrix A satisfies

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$$

P 32. Assume A is a stable square matrix and define the series

$$X \;\triangleq\; \sum_{n=0}^{\infty} A^n$$

(a) Show that the series converges.

(b) Show that $X = (I - A)^{-1}$.

P 33. Assume B is a stable square matrix and define the series

$$X \;\triangleq\; \sum_{n=0}^{\infty} [B^*]^n \, B^n$$

Show that the series converges to the unique solution of the Lyapunov equation $X - B^*XB = I$.

P 34. Establish Jensen's inequality for both cases of (F.26) and (F.29).

P 35. Establish both parts of Lemma F.5.

P 36. Establish Lemma F.6.

P 37. Derive the logit expression (G.6) from (G.5).

CHAPTER 2

P 38. Consider the gradient-descent recursion (2.40) where the step-size sequence is selected as

$$\mu(i) = \frac{\tau}{(i+1)^q}, \qquad \frac{1}{2} < q \leq 1, \quad \tau > 0$$

(a) Verify that the step-size sequence satisfies conditions (2.38).

(b) Determine the rate of convergence of $\|\widetilde{w}_i\|^2$ to zero.

(c) For a fixed τ, which value of q in the range $0.5 < q \leq 1$ results in the fastest convergence rate?

P 39. Consider the regularized logistic risk (2.9). Prove that

(a) $\|w^o\| \leq \mathbb{E}\,\|\boldsymbol{h}\|/\rho$.

(b) $\|w^o\|^2 \leq \text{Tr}(R_h)/\rho^2$.

P 40. Problems 40–44 are motivated by useful results from [190][Chs. 1,3]. Let $J(w)$ be a real-valued differentiable cost function whose gradient vector satisfies the Lipschitz condition (2.17). The cost $J(w)$ is not assumed be convex. Instead, we assume that it is lower-bounded, namely, $J(w) \geq L$ for all w and for some finite value L. Consider the gradient-descent algorithm (2.21). Show that if the step-size μ satisfies $\mu < 2/\delta$, then the sequence of iterates $\{w_i\}$ satisfy the following two properties:

(a) $J(w_i) \leq J(w_{i-1})$.

(b) $\lim_{i\to\infty} \nabla_w J(w_i) = 0$.

P 41. Let $J(w)$ be a real-valued cost function that satisfies the conditions stated in Assumption 2.1. Consider the gradient-descent algorithm (2.21). Establish the following result:

$$J(w_i) - J(w^o) \leq \;\alpha_1\,(J(w_{i-1}) - J(w^o))$$

where $\alpha_1 = 1 - 2\mu\nu + \mu^2\nu\delta$. Use this result to establish that $\widetilde{w}_i$ converges to zero for all $\mu < 2/\delta$ at a geometric rate determined by α_1. Compare this result with the statement of Lemma 2.1.

P 42. Refer to the second proof of Lemma 2.1. Conclude that convergence occurs at a geometric rate given by $\alpha_2 = \max\{(1-\mu\delta)^2, (1-\mu\nu)^2\}$. Show that the convergence rate is fastest when the step-size is chosen as $\mu^o = 2/(\nu+\delta)$ for which $\alpha_2^o = (\delta-\nu)^2/(\delta+\nu)^2$. Roughly, how many iterations are needed for the squared error, $\|\widetilde{w}_i\|^2$, to fall below a small threshold value, ϵ?

P 43. Let $J(w)$ denote a real-valued $\nu-$strongly convex and twice-differentiable cost function with $w \in \mathbb{R}^M$. Assume the Hessian matrix of $J(w)$ is $\delta-$Lipschitz continuous, i.e.,

$$\left\|\nabla_w^2 J(w_2) - \nabla_w^2 J(w_1)\right\| \;\leq\; \delta\|w_2 - w_1\|$$

The global minimizer of $J(w)$ is sought by means of the following iterative Newton's method:

$$w_i \;=\; w_{i-1} \;-\; \left[\nabla_w^2 J(w_{i-1})\right]^{-1} \nabla_{w^\mathsf{T}} J(w_{i-1}), \quad i \geq 0$$

which employs the inverse of the Hessian matrix. The initial condition is denoted by w_{-1}. Let

$$\alpha \overset{\Delta}{=} \left(\frac{\delta}{2\nu^2}\right)^2 \|\nabla_w J(w_{-1})\|^2$$

and assume $\alpha < 1$. Show that $\|\widetilde{w}_i\|^2$ converges to zero at a geometric rate. Specifically, show that

$$\|\widetilde{w}_i\|^2 \;\leq\; \left(\frac{2\nu^2}{\delta}\right)\alpha^{2^i}$$

Conclude that the convergence rate is now dependent on the quality of the initial condition.

P 44. Let $J(w)$ be a real-valued cost function that satisfies the conditions stated in Assumption 2.1 with $w \in \mathbb{R}^M$. Consider a modified gradient-descent algorithm of the following form:

$$w_i = w_{i-1} - \mu\nabla_{w^{\mathsf{T}}} J(w_{i-1}) \;+\; \eta(w_{i-1} - w_{i-2}), \quad i \geq 0$$

where the past iterate w_{i-2} is also used in the update equation. Assume the initial conditions w_{-1} and w_{-2} lie sufficiently close to w^o, i.e., $\|\widetilde{w}_{-1}\|^2 < \epsilon'$ and $\|\widetilde{w}_{-2}\|^2 < \epsilon'$ for some small enough ϵ.

(a) Show that if $0 \leq \eta < 1$ and $0 < \mu < 2(1+\eta)/\delta$, then $\|\widetilde{w}_i\|^2$ converges to zero at a geometric rate, α_3. Identify the rate and show that optimal values for $\{\mu, \eta, \alpha_3\}$ are

$$\mu^o = \frac{4}{\left(\sqrt{\delta}+\sqrt{\nu}\right)^2}, \qquad \eta^o = \left(\frac{\sqrt{\delta}-\sqrt{\nu}}{\sqrt{\delta}+\sqrt{\nu}}\right)^2, \qquad \alpha_3^o = \eta^o$$

(b) Let $\kappa = \delta/\nu$. Large values for κ indicate ill-conditioned Hessian matrices, $\nabla_w^2 J(w)$, since their spectra will lie over wider intervals. Let α_2^o denote the optimal rate of convergence when $\eta = 0$. We already know from Problem 42 that $\alpha_2^o = (\delta-\nu)^2/(\delta+\nu)^2$. Argue that for large κ:

$$\alpha_2^o \approx 1 - 2/\kappa, \qquad \alpha_3^o \approx 1 - 2/\sqrt{\kappa}$$

Compare the number of iterations that are needed for $\|\widetilde{w}_i\|^2$ to fall below a threshold ϵ for both cases of $\eta = 0$ and $\eta = \eta^o$.

P 45. Let $J(w)$ denote a real-valued $\nu-$strongly convex and twice-differentiable cost function with $w \in \mathbb{C}^M$. Following the construction from Problem 43, develop Newton's method for complex-valued arguments w, and study its convergence properties. Simplify the general results to the case of mean-square-error costs.

CHAPTER 3

P 46. How would the convergence rates shown in (3.111) change for step-size sequences of the form

$$\mu(i) = \frac{\tau}{(i+1)^q}, \qquad \frac{1}{2} < q \leq 1, \quad \tau > 0?$$

P 47. Assume the regression data $\boldsymbol{u}_i$ is Gaussian-distributed. Show that (3.22) is also satisfied for the following choice of the constant c:

$$c = \lambda^2_{\max}(R_u) \; + \; \lambda_{\max}(R_u)\mathrm{Tr}(R_u)$$

P 48. Let $\boldsymbol{\gamma}(i)$ be a streaming sequence of binary random variables that assume the values ± 1, and let $\boldsymbol{h}_i$ be a streaming sequence of $M \times 1$ real random (feature) vectors with $R_h = \mathbb{E}\,\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} > 0$. Assume the random processes $\{\boldsymbol{\gamma}(i), \boldsymbol{h}_i\}$ are wide-sense stationary. Consider the regularized logistic risk function:

$$J(w) \;=\; \frac{\rho}{2}\|w\|^2 \;+\; \mathbb{E}\left\{\ln\left[1 + e^{-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}w}\right]\right\}$$

(a) Write down the expression for the gradient noise process, $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, that would result from using a constant step-size stochastic gradient algorithm for seeking the minimum of $J(w)$.

(b) Verify that this noise process satisfies conditions similar to (3.31)–(3.32), namely,

$$\begin{aligned}
&\mathbb{E}\left[\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] = 0\\
&\mathbb{E}\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \leq \beta^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \sigma_s^2
\end{aligned}$$

for some nonnegative constants β^2 and σ_s^2.

(c) Verify similarly that the fourth-order moment of the noise process satisfies a condition similar to (3.56), namely,

$$\mathbb{E}\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \leq \beta_4^4\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; \sigma_{s4}^4$$

for some nonnegative constants β_4^4 and σ_{s4}^4. What conditions on the moments of the data are needed to ensure this result?

(d) For any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$, we let

$$R_{s,i}(\boldsymbol{w}) \;\stackrel{\Delta}{=}\; \mathbb{E}\left[\,\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right]$$

denote the conditional second-order moment of the gradient noise process. Show that the cost function, $J(w)$, and the above conditional moment satisfy Lipschitz conditions similar to (4.18)–(4.19), i.e.,

$$\begin{aligned}\left\|\nabla_w^2\, J(w^o+\Delta w)-\nabla_w^2\, J(w^o)\right\| &\leq \kappa_1\,\|\Delta w\| \\ \|R_{s,i}(w^o+\Delta w)-R_{s,i}(w^o)\| &\leq \kappa_2\,\|\Delta w\|^{\gamma}\end{aligned}$$

for small perturbations $\|\Delta w\|\leq\epsilon$ and for some constants $\kappa_1\geq 0$, $\kappa_2\geq 0$, and positive exponent γ. What conditions on the moments of the data are needed to ensure these results?

P 49. Consider the mean-square-error cost $J(w)=\mathbb{E}\,(\boldsymbol{d}(i)-\boldsymbol{u}_i w)^2$. Substitute (3.19) into (3.42) and verify that the error recursion in this case reduces to

$$\widetilde{\boldsymbol{w}}_i=(I_M-2\mu\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1}-2\mu\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i)$$

Refer to the conditions on the regression data and the measurement noise process $\{\boldsymbol{u}_i,\boldsymbol{v}(i)\}$ in Example 3.1. Assume further that the regression data is Gaussian distributed. The following problem is extracted from the results of [206][Ch. 23].

(a) Determine a recursion for $\mathbb{E}\,\widetilde{\boldsymbol{w}}_i$. Find a necessary and sufficient condition on μ to ensure convergence in the mean.

(b) Determine a recursion for $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$.

(c) Find a necessary and sufficient condition on μ to ensure that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ converges.

(d) How does the condition of part (c) compare with condition (3.36)?

(e) Find an expression for the MSD level of the filter.

P 50. Consider the mean-square-error cost $J(w)=\mathbb{E}\,(\boldsymbol{d}(i)-\boldsymbol{u}_i w)^2$. Refer to the conditions on the regression data and the measurement noise process $\{\boldsymbol{u}_i,\boldsymbol{v}(i)\}$ in Example 3.1. Assume further that the regression data is Gaussian distributed with $R_u=\sigma_u^2 I_M$. Consider the stochastic-gradient algorithm:

$$\boldsymbol{w}_i=\boldsymbol{w}_{i-1}+2\mu\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{e}(i),\quad \boldsymbol{e}(i)=\boldsymbol{d}(i)-\boldsymbol{u}_i\boldsymbol{w}_{i-1},\ \ i\geq 0$$

(a) Determine exact expressions for the filter EMSE and MSD, written as

$$\text{MSD}=\lim_{i\to\infty}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2,\qquad \text{EMSE}=\lim_{i\to\infty}\mathbb{E}\,|\boldsymbol{u}_i\widetilde{\boldsymbol{w}}_{i-1}|^2$$

where $\widetilde{\boldsymbol{w}}_i=w^o-\boldsymbol{w}_i$.

(b) Define the convergence time, $\mathcal{K}$, of the filter as the number of iterations it takes for the mean-square-error, $\mathbb{E}\,|\boldsymbol{e}(i)|^2$, to be within $\epsilon\%$ of its steady-state value. Find a closed form expression for $\mathcal{K}$.

P 51. Consider the LMS recursion (3.13) and a collection of L data points $\{\boldsymbol{d}(i), \boldsymbol{u}_i,\; i = 0, 1, \ldots, L-1\}$. Starting with the initial condition $\boldsymbol{w}_{-1} = 0$, the recursion is applied to all L data points until $\boldsymbol{w}_{L-1}$ is generated. The recursion is then applied again to the same data points albeit using now $\boldsymbol{w}_{L-1}$ as the initial condition. This process is continued indefinitely: at the end of every L iterations, the procedure is repeated starting from the last iterate obtained in the previous iteration. Assuming the step-size μ is sufficiently small and L is finite but of sufficient size, what would the MSD performance of this implementation be?

P 52. Refer to the stochastic gradient recursion (3.5) and assume the step-size parameter μ is replaced by a diagonal matrix as follows:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - B\,\widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0$$

where B is the $M \times M$ diagonal matrix:

$$B \;\overset{\Delta}{=}\; \mu_{\max} \cdot \text{diag}\{b_1, b_2, \ldots, b_M\}$$

and the $\{b_m\}$ are positive scalars in the range $0 < b_m \leq 1$. In other words, different step-sizes are possibly assigned to the different entries of $\boldsymbol{w}_i$, with $\mu_{\max}$ denoting the largest step-size value. Assume the conditions under Assumptions 3.1 and 3.2 on the cost function and the gradient noise process continue to hold.

(a) Extend the result of Lemma 3.1 to this case.

(b) Extend the result of Lemma 3.2 to this case.

P 53. All variables are zero-mean. Consider a complex-valued scalar random variable $\boldsymbol{d}$ and a complex-valued $1 \times M$ regression vector $\boldsymbol{u}$. Let $\widehat{\boldsymbol{d}} = \boldsymbol{u}w^o$ denote the linear least-mean-squares estimator of $\boldsymbol{d}$ given $\boldsymbol{u}$. That is, $w^o \in \mathbb{C}^M$ is the vector that minimizes the mean-square-error cost

$$w^o \;\overset{\Delta}{=}\; \arg\min_{w} \mathbb{E}\,|\boldsymbol{d} - \boldsymbol{u}w|^2$$

Consider additionally the problems of estimating separately the real and imaginary parts of $\boldsymbol{d}$ from the real and imaginary parts of $\boldsymbol{u}$, also in the linear least-mean-squares error sense, namely,

$$\widehat{\boldsymbol{d}}_{\text{real}} = \begin{bmatrix} \text{Re}(\boldsymbol{u}) & \text{Im}(\boldsymbol{u}) \end{bmatrix} w^o_{\text{real}}, \quad \widehat{\boldsymbol{d}}_{\text{imag}} = \begin{bmatrix} \text{Re}(\boldsymbol{u}) & \text{Im}(\boldsymbol{u}) \end{bmatrix} w^o_{\text{imag}}$$

where the $w^o_{\text{real}} \in \mathbb{R}^{2M}$ and $w^o_{\text{imag}} \in \mathbb{R}^{2M}$ are the minimizers to the following mean-square-error costs:

$$\begin{aligned} w^o_{\text{real}} &\overset{\Delta}{=} \arg\min_{w_{\text{real}}} \mathbb{E} \left| \text{Re}(\boldsymbol{d}) \; - \left[\; \text{Re}(\boldsymbol{u}) \quad \text{Im}(\boldsymbol{u}) \; \right] w_{\text{real}} \right|^2 \\ w^o_{\text{imag}} &\overset{\Delta}{=} \arg\min_{w_{\text{imag}}} \mathbb{E} \left| \text{Im}(\boldsymbol{d}) - \left[\; \text{Re}(\boldsymbol{u}) \quad \text{Im}(\boldsymbol{u}) \; \right] w_{\text{imag}} \right|^2 \end{aligned}$$

Let $\widehat{\boldsymbol{d}}_2 = \widehat{\boldsymbol{d}}_{\text{real}} + j\widehat{\boldsymbol{d}}_{\text{imag}}$ denote the estimator that is obtained for $\boldsymbol{d}$ from the second construction.

(a) Argue that the problem of estimating the real and imaginary parts of $\boldsymbol{d}$ from the real and imaginary parts of $\boldsymbol{u}$ is equivalent to the problem of estimating $\boldsymbol{d}$ from the combination $\{\boldsymbol{u}, \boldsymbol{u}^*\}$, namely,

$$\widehat{\boldsymbol{d}}_2 = \boldsymbol{u}a^o \; + \; (\boldsymbol{u}^*)^{\mathsf{T}}\, b^o$$

where $a^o, b^o \in \mathbb{C}^M$ correspond to the minimizer of the mean-square-error cost:

$$a^o, b^o \overset{\Delta}{=} \arg\min_{a,b} \mathbb{E}\, |\boldsymbol{d} - \boldsymbol{u}a - (\boldsymbol{u}^*)^{\mathsf{T}} b|^2$$

(b) Determine expressions for w^o, w^o_{real}, and w^o_{imag}.

(c) What is the mean-square-error that results from the construction $\widehat{\boldsymbol{d}}_2$? How does it compare to the mean-square-error obtained from the construction $\widehat{\boldsymbol{d}}$? Under what conditions will both constructions lead to the same mean-square-error value?

(d) Write down an LMS-type stochastic-gradient algorithm for estimating $\{a^o, b^o\}$ from streaming data $\{\boldsymbol{d}(i), \boldsymbol{u}_i, \boldsymbol{u}_i^*\}$.

(e) Assuming sufficiently small step-sizes, derive an expression for the filter MSD. How does this performance compare to that delivered by the traditional LMS implementation that estimates w^o directly from $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$? Under what conditions will both filters have similar MSD performance?

CHAPTER 4

P 54. Refer to the stochastic gradient recursion (3.5) and assume the step-size parameter μ is replaced by a diagonal matrix as follows:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - B\, \widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0$$

where B is the $M \times M$ diagonal matrix:

$$B \overset{\Delta}{=} \mu_{\max} \cdot \text{diag}\{b_1, b_2, \ldots, b_M\}$$

and the $\{b_m\}$ are positive scalars in the range $0 < b_m \leq 1$. In other words, different step-sizes are possibly assigned to update the different entries of $\boldsymbol{w}_i$, with $\mu_{\max}$ denoting the largest step-size value. Assume the conditions under Assumptions 4.1, 4.2, and 4.4 on the cost function and the gradient noise process continue to hold. Extend the result of Theorem 4.7 to this case. What would the performance expressions be in the complex case?

P 55. Is there any advantage to using a complex step-size in the complex case for single-agent adaptation and learning? Refer to the stochastic-gradient recursion (3.116) in the complex case and replace μ by the complex value $\mu = \mu_a e^{j\theta}$, where μ_a denotes its amplitude and θ denotes its phase. How would the results of Lemma 3.5 and Theorem 4.8 be modified?

CHAPTER 5

P 56. Consider a collection of N agents, each running an LMS update with step-size μ_k similar to the situation described in Section 5.1. We express the step-sizes across the agents in the form $\mu_k = \mu_{\max} \cdot b_k$, where b_k is a positive scalar bounded by one. What is the average non-cooperative performance in this case? At what rate of convergence will this average performance be achieved? Next consider a centralized implementation of the form (5.13). Pick μ to match the above convergence rate. How do the performance levels of the average non-cooperative and centralized solutions compare in this case? Will the conclusion of Lemma 5.2 continue to hold?

P 57. Lemma 5.2 establishes that the MSD performance of the centralized stochastic-gradient solution is always superior to the average MSD performance over a collection of N non-cooperative agents. Example 5.1 describes a situation where the centralized solution is $N-$fold superior to the non-cooperative solution. Can you find an example where the centralized solution can be more than $N-$fold superior to non-cooperative implementations?

CHAPTER 6

P 58. Is the converse statement of Lemma 6.1 correct? That is, does it hold that if A is a primitive left-stochastic matrix, then the network is strongly connected?

P 59. Consider an $N-$agent connected network with a left-stochastic combination matrix A. We already know that A is irreducible but not necessarily primitive. Let $B = 0.5(I + A)$. Is B a left-stochastic matrix? Show that the entries of B^{N-1} are all positive. Conclude that B is a primitive matrix.

P 60. Show that to check whether an $N \times N$ left-stochastic matrix A is irreducible or primitive, we can replace all nonzero entries in A by ones and verify instead whether the resulting matrix is irreducible or primitive.

P 61. Assume A is a left-stochastic primitive matrix of size $N \times N$.

(a) Show that A is power convergent and the limit converges to the following rank-one product:

$$\lim_{n\to\infty} A^n \;=\; p\mathbb{1}^{\mathsf{T}}$$

where p is the Perron vector of A. Is the limit a primitive matrix?

(b) For any vector $b = \text{col}\{b_1, b_2, \ldots, b_N\}$, show that

$$\lim_{n\to\infty} A^n b \;=\; \alpha p$$

where $\alpha = b_1 + b_2 + \ldots + b_N$.

(c) If A is irreducible but not necessarily primitive, does the limit of part (a) exist?

P 62. Consider the 3×3 combination matrix

$$A = \begin{bmatrix} \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}$$

Draw the corresponding graph. Is the network strongly-connected? Compute the powers A^3 and A^4. Conclude that the power n_o in (6.10) can be larger than the number of agents, N.

P 63. Give an example of a $4-$agent network that is connected (but not strongly-connected) and whose combination matrix A is not primitive. Verify that the corresponding A is indeed not primitive.

P 64. Consider a strongly-connected network with $N-$agents. Prove that, for any agent k, there always exists a circular path (i.e., a cycle) with non-zero scaling weights that starts at k and ends at the same location k.

P 65. Show that a network is connected if, and only if, its combination matrix A is irreducible.

P 66. Consider an $N \times N$ left-stochastic matrix A. Let $n_o = N^2 - 2N + 2$. Show that A is primitive if, and only if, $[A^{n_o}]_{\ell k} > 0$ for all ℓ and k.

P 67. Refer to the original consensus recursion (7.35). Show that under condition (7.33) and assuming that $|\lambda_2(A)| < 1$, the convergence result (7.36) holds for all agents. Determine the rate of convergence.

P 68. Consider a network consisting of N vertices and L edges. We associate two useful matrices with the network. One is the Laplacian matrix, denoted by $\mathcal{L}$, and is square of size $N \times N$, while the second matrix is the incidence matrix, denoted by $\mathcal{I}$, and its size is size $N \times L$. The Laplacian matrix is useful in characterizing whether a network consists of a single graph or of separate disconnected subgraphs.

We denote the degree of an agent k by n_k and define it as the size of its neighborhood, $n_k = |\mathcal{N}_k|$. Since $k \in \mathcal{N}_k$, we have $n_k \geq 1$. The Laplacian matrix is symmetric and its entries are defined as follows:

$$[\mathcal{L}]_{k\ell} = \begin{cases} n_k - 1, & \text{if } k = \ell \\ -1, & \text{if } k \neq \ell \text{ and nodes } k \text{ and } \ell \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases}$$

Note that the term $n_k - 1$ measures the number of edges that are incident on agent k, and the locations of the $-1's$ on row k indicate the agents that are connected to agent k. The entries of the incidence matrix, $\mathcal{I}$, are defined as follows. Every column of $\mathcal{I}$ represents one edge in the graph. Each edge connects two agents and its column will display two nonzero entries at the rows corresponding to these agents: one entry will be $+1$ and the other entry will be -1. For directed graphs, the choice of which entry is positive or negative can be used to identify the agents from which edges emanate (source nodes) and the agents at which edges arrive (sink nodes). We shall simply assign positive values to lower indexed agents and negative values to higher indexed agents:

$$[\mathcal{I}]_{ke} = \begin{cases} +1, & \text{if agent } k \text{ is the lower-indexed node connected to edge } e \\ -1, & \text{if agent } k \text{ is the higher-indexed node connected to edge } e \\ 0, & \text{otherwise} \end{cases}$$

Let

$$\theta_1 \geq \theta_2 \geq \ldots \geq \theta_N$$

denote the ordered eigenvalues of $\mathcal{L}$. Establish the following properties — See App. B of [208]:

(a) $\mathcal{L} = \mathcal{I}\ \mathcal{I}^{\mathsf{T}}$. Conclude that $\mathcal{L}$ is symmetric nonnegative-definite so that $\theta_m \geq 0$.

(b) The rows of $\mathcal{L}$ add up to zero so that $\mathcal{L}\mathbb{1} = 0$. This means that $\mathbb{1}$ is a right eigenvector of $\mathcal{L}$ corresponding to the eigenvalue zero.

(c) The smallest eigenvalue is always zero, $\theta_N = 0$. The second smallest eigenvalue, θ_{N-1}, is called the algebraic connectivity of the graph.

(d) The number of times that zero is an eigenvalue of $\mathcal{L}$ (i.e., its multiplicity) is equal to the number of connected subgraphs.

(e) The algebraic connectivity of a connected graph is nonzero, i.e., $\theta_{N-1} \neq 0$. In other words, a graph is connected if, and only if, its algebraic connectivity is nonzero.

P 69. Show that the Laplacian matrix of a fully-connected network with N agents is given by $\mathcal{L} = NI_N - \mathbb{1}_N\mathbb{1}_N^{\mathsf{T}}$. Conclude that the largest eigenvalue of $\mathcal{L}$ is $\lambda_1(\mathcal{L}) = N$.

CHAPTER 7

P 70. Write down the ATC diffusion LMS algorithm that would result from minimizing the following regularized cost function over a connected network of N agents, where $\rho > 0$ and $w \in \mathbb{C}^M$:

$$\min_w \quad \rho\|w\|^2 \;+\; \sum_{k=1}^{N} \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2$$

P 71. Derive a CTA diffusion LMS algorithm for minimizing the following constrained cost function over a connected network of N agents:

$$\min_w \sum_{k=1}^{N} \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2 \quad \text{subject to} \quad c^*w = \alpha$$

where c is a known column vector and α is a given scalar.

P 72. Derive a consensus LMS algorithm for minimizing the following constrained cost function over a connected network of N agents:

$$\min_w \sum_{k=1}^{N} \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2 \quad \text{subject to} \quad c^*w = \alpha$$

where c is a known column vector and α is a given scalar.

P 73. Derive a diffusion least-mean-squares algorithm for minimizing the following optimization problem over a connected network of $N = N_1 + N_2$ agents:

$$\min_w \left[\sum_{k=1}^{N_1} \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2 \;-\; \sum_{k=N_1+1}^{N_1+N_2} \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2 \right]$$

Can you provide an interpretation for this choice of cost functions?

CHAPTER 8

P 74. Consider the incremental LMS recursion (7.7). Define the network error vector at time i:

$$\widetilde{\boldsymbol{w}}_i \;\overset{\Delta}{=}\; \text{col}\,\{\widetilde{\boldsymbol{w}}_{1,i}, \widetilde{\boldsymbol{w}}_{2,i}, \ldots, \widetilde{\boldsymbol{w}}_{N,i}\}$$

Determine a recursion for the evolution of $\widetilde{\boldsymbol{w}}_i$.

P 75. Consider the diffusion logistic recursion (7.26). Define the network error vector at time i:

$$\widetilde{\boldsymbol{w}}_i \;\overset{\Delta}{=}\; \text{col}\,\{\widetilde{\boldsymbol{w}}_{1,i}, \widetilde{\boldsymbol{w}}_{2,i}, \ldots, \widetilde{\boldsymbol{w}}_{N,i}\}$$

Determine a recursion for the evolution of $\widetilde{\boldsymbol{w}}_i$.

P 76. Consider consensus and diffusion strategies with enlarged cooperation schemes similar to recursions (7.27) and (7.28), which involve a right-stochastic matrix C. Repeat the derivation of Example 8.1 for MSE networks and derive the analogue of recursions (8.22) and (8.25) when C is present.

P 77. Consider a situation similar to the MSE network model described in Example 6.3, except that the linear regression model at each agent k is now

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i}w_k^o + \boldsymbol{v}_k(i), \quad k = 1, 2, \ldots, N$$

where the models $\{w_k^o\}$ are possibly different across the agents. For each agent k, we define the error vector $\widetilde{\boldsymbol{w}}_{k,i} = w_k^o - \boldsymbol{w}_{k,i}$, where the error is measured relative to w_k^o. Let

$$\widetilde{\boldsymbol{w}}_i \;\overset{\Delta}{=}\; \text{col}\{\widetilde{\boldsymbol{w}}_{1,i}, \widetilde{\boldsymbol{w}}_{2,i}, \ldots, \widetilde{\boldsymbol{w}}_{N,i}\}$$

Repeat the arguments of Example 8.1 and derive the corresponding error recursions that would correspond to the extensions of (8.22) and (8.25) to this scenario.

P 78. Establish the validity of (8.91), which provides an expression for the entries of the Perron eigenvector of a uniform combination matrix.

P 79. Show that (8.98) are indeed the entries of the Perron eigenvector of the Hastings matrix defined by (8.96).

P 80. Refer to Figure 8.2. Argue, from the definition of Pareto optimality, that any $w \in (w_1^o, w_2^o)$ is a Pareto optimal solution for the multi-objective optimization problem

$$\min_{w} \{J_1(w)\, J_2(w)\}$$

where $J_1(w)$ and $J_2(w)$ are the quadratic costs defined by (8.69). How would you select the scalarization weights $\{\pi_1, \pi_2\}$ to ensure that the resulting Pareto optimal solution is $w^\pi = \frac{1}{3}w_1^o + \frac{2}{3}w_2^o$? Likewise, how would you select the scalarization weights $\{\pi_1, \pi_2\}$ to ensure that the resulting Pareto optimal solution is $w^\pi = \frac{2}{3}w_1^o + \frac{1}{3}w_2^o$?

P 81. Pareto optimality is a useful concept in game theory as well. Consider two players, A and B. Each player has a choice between two strategies: Player A can choose between strategies A1 or A2, while player B can choose between strategies B1 or B2. The table below lists the costs associated with the four possible choices by the players. For example, refer to the entry in the table corresponding to player A selecting A1 and player B selecting B1. The cell shows the values $(6, 4)$, meaning that 6 is the cost incurred by player A and 4 is the cost incurred by player B. The players wish to minimize their costs. Can you identify which strategies are Pareto optimal? That is, can you identify those strategies such that there are no other strategies where at least one player sees his cost reduced (i.e., does better) while the other player does not do worse?

	B1	B2
A1	(6,4)	(5,5)
A2	(4,6)	(7,5)

P 82. Assume all agents employ the same step-size parameter, $\mu_k \equiv \mu$. Assume further that we would like a strongly-connected network of N agents to optimize the weighted aggregate cost:

$$J^{\text{glob},\pi}(w) \triangleq \sum_{k=1}^{N} \pi_k J_k(w)$$

(a) Consider initially the case in which the positive weight π_k that is associated with agent k is chosen to be proportional to its relative degree (i.e., to its level of connectivity in the network), namely,

$$\pi_k \triangleq n_k \left(\sum_{\ell=1}^{N} n_\ell\right)^{-1}, \quad n_k \triangleq |\mathcal{N}_k|$$

What combination policy would result from the construction (8.96)? Is the policy left-stochastic or doubly-stochastic?

(b) Consider next the case in which all π_k are identical so that all individual costs are weighted equally, and minimizing $J^{\text{glob},\pi}(w)$ is equivalent to minimizing the aggregate cost:

$$J^{\text{glob}}(w) \;\overset{\Delta}{=}\; \sum_{k=1}^{N} J_k(w)$$

What combination policy would result from the construction (8.96)? Is it left-stochastic or doubly-stochastic?

P 83. If each of the individual left-stochastic combination matrices $\{A_o, A_1, A_2\}$ is primitive, is the product $P = A_1 A_o A_2$ defined by (8.48) primitive? Conversely, if P is primitive, does it necessarily follow that each of the matrices $\{A_o, A_1, A_2\}$ is primitive? Prove or give a counter-example.

P 84. Refer to the general strategy (8.46) and observe that each line contains a sum over agents ℓ in a neighborhood $\mathcal{N}_k$; these sums use combination weights from the matrices $\{A_1, A_o, A_2\}$. Do these neighborhoods need to coincide? If we write instead

$$\left\{\begin{array}{lcl}
\boldsymbol{\phi}_{k,i-1} & = & \displaystyle\sum_{\ell\in\mathcal{N}_{k,1}} a_{1,\ell k}\, \boldsymbol{w}_{\ell,i-1} \\
\boldsymbol{\psi}_{k,i} & = & \displaystyle\sum_{\ell\in\mathcal{N}_{k,o}} a_{o,\ell k}\, \boldsymbol{\phi}_{\ell,i-1} \;-\; \mu_k\, \widehat{\nabla_{w^*} J}_k\left(\boldsymbol{\phi}_{k,i-1}\right) \\
\boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell\in\mathcal{N}_{k,2}} a_{2,\ell k}\, \boldsymbol{\psi}_{\ell,i}
\end{array}\right.$$

where $\{\mathcal{N}_{k,1}, \mathcal{N}_{k,o}, \mathcal{N}_{k,2}\}$ refer to neighborhoods defined by the respective matrices $\{A_1, A_o, A_2\}$, would the result of Lemma 8.1 still hold?

P 85. Refer to the general strategy (8.46) and replace it by

$$\left\{\begin{array}{lcl}
\boldsymbol{\phi}_{k,i-1} & = & \displaystyle\sum_{\ell\in\mathcal{N}_{k,1}} a_{1,\ell k}\, \boldsymbol{w}_{\ell,i-1} \\
\boldsymbol{\psi}_{k,i} & = & \displaystyle\sum_{\ell\in\mathcal{N}_{k,o}} a_{o,\ell k}\, \boldsymbol{\phi}_{\ell,i-1} \;-\; \mu_k \sum_{\ell\in\mathcal{N}_{k,c}} c_{\ell k}\, \widehat{\nabla_{w^*} J}_\ell\left(\boldsymbol{\phi}_{k,i-1}\right) \\
\boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell\in\mathcal{N}_{k,2}} a_{2,\ell k}\, \boldsymbol{\psi}_{\ell,i}
\end{array}\right.$$

where $C = [c_{\ell k}]$ is a right-stochastic matrix (each of its rows adds up to one), and $\{\mathcal{N}_{k,1}, \mathcal{N}_{k,o}, \mathcal{N}_{k,2}, \mathcal{N}_{k,c}\}$ refer to neighborhoods defined by the respective matrices $\{A_1, A_o, A_2, C\}$. Extend the result of Lemma 8.1 to this case and derive the corresponding error recursion.

CHAPTER 9

P 86. Refer to definition (9.7) for q and let q_k denote the individual entries of q. Show that

$$\sum_{k=1}^{N} \frac{q_k}{\mu_k} = 1$$

P 87. Refer to the variable $\bar{\boldsymbol{w}}_i^e$ defined by (9.55). What are the dimensions of $\bar{\boldsymbol{w}}_i^e$? Provide an interpretation for $\bar{\boldsymbol{w}}_i^e$ as a weighted linear combination of the error vectors $\{\widetilde{\boldsymbol{w}}_{k,i}^e\}$ across all N agents.

P 88. Refer to the non-cooperative strategy (5.76) with step-size μ_k, say,

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \, \widehat{\nabla_{w^*} J}_k(\boldsymbol{w}_{k,i-1}), \quad k = 1, 2, \ldots, N$$

and introduce the weighted convex combination

$$\boldsymbol{w}_{c,i} \;\overset{\Delta}{=}\; \sum_{k=1}^{N} p_k \boldsymbol{w}_{k,i}$$

for some positive scalars $\{p_k\}$ that add up to one.

(a) Apply the mean-value relation (D.20) around the minimizer $w^\star$ of (9.6) and derive a recursion for the extended version of the error vector $\widetilde{\boldsymbol{w}}_{c,i} = w^\star - \boldsymbol{w}_{c,i}$. How does the recursion for $\widetilde{\boldsymbol{w}}_{c,i}^e$ compare with the recursion for $\bar{\boldsymbol{w}}_i^e$ given by (9.60)?

(b) Evaluate the order of the following mean-square-error:

$$\limsup_{i\to\infty} \; \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i} - \widetilde{\boldsymbol{w}}_{c,i}\|^2\,?$$

What is the interpretation of this result?

P 89. Consider the case of MSE networks from Example 6.3 with a common minimizer w^o across all agents, and assume the ATC diffusion strategy is employed by setting $A_1 = A_o = I_N$ and $A_2 = A$ in (8.46) for some left-stochastic primitive matrix A. Repeat the proof of Theorem 9.1 for this case, with particular attention to the steps that get simplified. Are extended error vectors necessary in this case?

P 90. Consider the case of MSE networks from Example 6.3 with a common minimizer w^o across all agents, and assume the ATC diffusion strategy is employed by setting $A_1 = A_o = I_N$ and $A_2 = A$ in (8.46) for some doubly-stochastic primitive matrix A. Write down what the corresponding error recursion (9.12) will become in this case. Are extended error vectors necessary in this case? Equate the variances of both sides of this error recursion and use the result to establish the mean-square stability of the network without the need to introduce the Jordan canonical decomposition, and the basis transformation, used in the proof of Theorem 9.1

P 91. Refer to the general strategy (8.46) and replace it by

$$\left\{\begin{array}{lcl} \boldsymbol{\phi}_{k,i-1} & = & \displaystyle\sum_{\ell \in \mathcal{N}_{k,1}} a_{1,\ell k}\, \boldsymbol{w}_{\ell,i-1} \\ \boldsymbol{\psi}_{k,i} & = & \displaystyle\sum_{\ell \in \mathcal{N}_{k,o}} a_{o,\ell k}\, \boldsymbol{\phi}_{\ell,i-1} \;-\; \mu_k \sum_{\ell \in \mathcal{N}_{k,c}} c_{\ell k} \widehat{\nabla_{w^*} J}_\ell \left(\boldsymbol{\phi}_{k,i-1}\right) \\ \boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell \in \mathcal{N}_{k,2}} a_{2,\ell k}\, \boldsymbol{\psi}_{\ell,i} \end{array}\right.$$

where $C = [c_{\ell k}]$ is a right-stochastic matrix (each of its rows adds up to one), and $\{\mathcal{N}_{k,1}, \mathcal{N}_{k,o}, \mathcal{N}_{k,2}, \mathcal{N}_{k,c}\}$ refer to neighborhoods defined by the respective matrices $\{A_1, A_o, A_2, C\}$. Does the result of Theorem 9.1 still hold? What steps in the derivation will be affected?

P 92. Assume the matrix P corresponds to a connected but not necessarily strongly-connected network (i.e., P is not necessarily primitive any longer). Refer to Theorem 9.1. Can we still ensure the mean-square-error stability of the network? Prove or give a counter-example.

P 93. Refer to the expressions for $\mathcal{B}_{\rm atc}$ and $\mathcal{B}_{\rm cta}$ in (9.174)–(9.175). Assume the non-cooperative matrix $\mathcal{B}_{\rm ncop} = I_{2MN} - \mathcal{M}\mathcal{H}$ is stable. Show that $\mathcal{B}_{\rm atc}$ and $\mathcal{B}_{\rm cta}$ are stable regardless of whether the combination matrix A is primitive or not, and regardless of whether the network topology is connected or not.

P 94. Simplify expression (9.173) for the case of MSE networks studied in Example 6.3.

P 95. Verify the validity of the right-hand side of (9.214).

P 96. Consider the setting of Theorem 9.1 and assume the ATC diffusion strategy is employed by setting $A_1 = A_o = I_N$ and $A_2 = A$ in (8.46) for some doubly-stochastic primitive matrix A. Write down what the corresponding error recursion (9.12) will become in this case. Equate the variances of both sides of this error recursion and use the result to establish the mean-square stability of the network without the need to introduce the Jordan canonical decomposition, and the basis transformation, used in the proof of Theorem 9.1. Can you repeat the same argument if CTA diffusion is used instead? What about consensus?

CHAPTER 10

P 97. How would expressions (10.29) and (10.30) be modified for the case of MSE networks from Example 6.3?

P 98. Refer to Example 10.1 and assume all agents employ the same step-size, $\mu_k \equiv \mu$, and that the Hessian matrices are uniform across the agents, $H_k \equiv H$. For example, this scenario arises for MSE networks of the form studied in Example 6.3 when all agents employ the same step-size and observe regression data with uniform covariance matrix, R_u. Show that in this case:

$$\rho(\mathcal{B}_{\text{diff}}) \;=\; \rho(\mathcal{B}_{\text{ncop}}) \;\leq\; \rho(\mathcal{B}_{\text{cons}})$$

with equality holding when $A = I_N$ or when the step-size satisfies

$$0 < \mu < \min_{m \neq 1} \left\{ \frac{1 - |\lambda_m(A)|}{\lambda_{\min}(H) + \lambda_{\max}(H)} \right\}$$

where $\lambda_1(A) = 1$. What is the interpretation of this result?

P 99. Assume the mean-error recursion for some non-cooperative agents is unstable and that the combination matrix, A, is symmetric. Can the consensus strategy stabilize the network for any A?

P 100. Even if the mean-error recursion for some non-cooperative agents is unstable, the diffusion strategy can still stabilize the network. Why?

P 101. True of False: For an $N-$agent MSE network with a symmetric combination policy, if the step-sizes are fixed and at least one non-cooperative agent is unstable in the mean, then the consensus network is unstable regardless of the topology. Prove or give a counter-example.

P 102. What would the error recursion (10.13) reduce to in the case of the MSE networks described in Example 6.3? What is the value of $\boldsymbol{c}_{i-1}$ in that case?

P 103. What would the long-term model (10.19) be in the case of the MSE networks described in Example 6.3. How is this model different from the original recursion (10.13)?

P 104. Consider the specialization of Example 10.1 to an MSE network (of the form described in Example 6.3) running on real data. Give an example of a strongly-connected network with $N = 4$ agents such that (a) the non-cooperative strategy is stable in the mean; (b) the ATC and CTA diffusion strategies are stable in the mean; and (c) the consensus strategy is unstable in the mean for the same step-size and combination policy as the diffusion strategies.

CHAPTER 11

P 105. Can you minimize the MSD expression (11.144) over the $\{p_k\}$? Can you maximize the same MSD expression over the $\{p_k\}$?

P 106. How does result (11.144) compare with the average performance of $N-$non-cooperative agents? When will it be smaller?

P 107. How does result (11.203) compare with the average performance of $N-$non-cooperative agents? When will it be smaller?

P 108. How does result (11.203) compare with the performance of the centralized solution given by (11.204)?

P 109. How does result (11.151) compare with the average performance of $N-$non-cooperative agents? When will it be smaller?

P 110. Establish the validity of (11.240) and (11.242).

P 111. Derive expressions (11.247)–(11.250).

P 112. Establish result (11.251).

P 113. Derive expressions (11.235)–(11.238).

P 114. Refer to the simulation in Figure 10.2. Use expression (11.156) to estimate what the steady-state MSD value should be. Does this value match well with the limiting value of the learning curves in the figure? Use instead expression (11.178) to estimate the steady-state MSD value. How does this value now compare to the limiting value of the learning curves in the figure? How do you explain the discrepancy? Can you derive a better expression for the MSD in this case?

CHAPTER 12

P 115. Establish the validity of the algebraic property (12.31).

P 116. Is the Hastings matrix A^o defined by (12.20) the only solution to (12.18)? Can you find another solution?

P 117. Refer to the Hastings matrix A^o defined by (12.20). When does this rule reduce to the averaging rule defined by (8.89)?

P 118. Refer to the MSD expression (12.5) for a distributed solution. Simplify the expression for both cases when the combination policy A is the Metropolis rule and when the combination policy A is the averaging rule.

(a) Can you compare these two MSD values directly?

(b) Give examples showing when one rule outperforms the other.

(c) Give a condition that ensures equal MSD values for both rules.

(d) Under what conditions on the network topology, does the averaging rule outperform the Metropolis rule?

(e) Under what conditions on the network topology, does the Metropolis rule outperform the averaging rule?

P 119. Refer to the MSD expression (12.5) for a distributed solution. Can you derive a sufficient condition under which a left-stochastic combination policy would outperform a doubly-stochastic policy?

P 120. Let w^o denote the unique minimizer of the aggregate cost

$$J^{\rm glob}(w) = \sum_{k=1}^{N} J_k(w)$$

and consider a weighted centralized solution of the form:

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\left(\sum_{k=1}^{N} p_k\, \widehat{\nabla_{w^*} J}_k(\boldsymbol{w}_{i-1})\right), \quad i \geq 0$$

for some positive coefficients $\{p_k\}$ that add up to one.

(a) Determine an expression for the MSD of this solution in terms of the $\{p_k\}$ under the same conditions on the aggregate cost function and the gradient noise process as those stated in Theorem 5.1.

(b) Assume all individual costs have uniform Hessian matrices $H_k = \nabla_w^2 J_k(w^o)$, i.e., $H_k \equiv H$ for $k = 1, 2, \ldots, N$. Optimize the MSD expression over the $\{p_k\}$ and show that their optimal values are again given by (12.21).

CHAPTER 13

P 121. Refer to the MSD expression (13.24) for the distributed solution with informed agents. If we fix the size of $\mathcal{N}_I$, there are many subgroups of size N_I that can be chosen to serve as informed agents. How many? If we were to pick which subgroup results in the smallest MSD value, how would you do that? Does this subgroup also result in the fastest convergence rate according to (13.23). Either solve analytically or provide examples with constructions that illustrate your answers.

P 122. Start with a strongly-connected network with N agents and assume we number the agents such that the resulting Perron entries $\{p_k\}$ are ordered from largest to smallest, i.e., $p_1 \geq p_2 \geq \ldots \geq p_N > 0$. Assume further that the Hessian matrices are uniform across the agents, i.e., $H_k \equiv H$ for $k = 1, 2, \ldots, N$. If we were to pick a subset of informed agents of size N_I to ensure fastest convergence rate according to (13.23), which agents would you pick? Does this selection result in the smallest MSD value according to (13.24) over all possible selections of N_I informed agents? Either solve analytically or provide examples with constructions that illustrate your answers.

P 123. Consider the setting of Example 13.1. Assume the network employs the averaging combination rule and that all agents have the same uniform degree, n. That is, $|\mathcal{N}_k| = n$ for all $k = 1, 2, \ldots, N$. If we were to pick a subset of informed agents of size N_I to ensure fastest convergence rate according to (13.35), does it matter which agents we pick? Which selection of agents results in the smallest MSD value according to (13.36) over all possible subsets of N_I informed agents?

P 124. Refer to expression (9.173) for $\rho(\mathcal{B})$ and consider the approximation

$$\rho(\mathcal{B}) \;\approx\; 1 - \lambda_{\min}\left(\sum_{k=1}^{N} q_k H_k\right)$$

Assume further that each individual Hessian matrix, H_k, is positive-definite and independent of $w^\star$. For example, this situation arises for the class MSE networks described in Example 6.3.

(a) Assume initially that $H_k \equiv H > 0$ for all k. Can you optimize $\rho(\mathcal{B})$ over the $\{q_k\}$ so as to result in a spectral radius that is the furthest from one possible? Which combination policy A would achieve this optimal rate of convergence?

(b) Assume now that $H_k > 0$ but that these Hessian matrices are not necessarily uniform across the agents. Can you again optimize $\rho(\mathcal{B})$ over the $\{q_k\}$ so as to result in a spectral radius that is the furthest from one possible? Which combination policy A would achieve this optimal rate of convergence?

P 125. Establish conclusion (13.47).

CHAPTER 14

P 126. Verify that the Laplacian rule in Table 14.1 is symmetric and doubly-stochastic.

P 127. Verify that the relative-degree rule in Table 14.1 is left-stochastic.

P 128. Establish result (14.22).

P 129. Establish the validity of (14.27).

P 130. Refer to listings (14.51) and (14.57) for the computation of the agent-centered and neighbor-centered adaptive combination weights. Repeat the arguments leading to these listings to justify that for the CTA diffusion strategy (7.18), the vectors $\boldsymbol{y}_{\ell,i}$ in (14.51) and $\boldsymbol{y}_{\ell k,i}$ in (14.57) can be defined as follows:

$$\boldsymbol{y}_{\ell,i} \;\; \triangleq \;\; \boldsymbol{w}_{\ell,i} - \boldsymbol{\psi}_{\ell,i-1}, \qquad \boldsymbol{y}_{\ell k,i} \;\; \triangleq \;\; \boldsymbol{w}_{k,i} - \boldsymbol{w}_{\ell,i-1}$$

CHAPTER 15

P 131. Refer to the gossip strategy (15.1) and assume each agent k selects ℓ_o uniformly from among its neighbors, i.e., with probability $1/(n_k-1)$. Assume the combination coefficient a_k is uniform across all agents and denote it by $a \in [0,1]$. Assume complex-valued data. Derive an expression for the network MSD for sufficiently small step-sizes.

P 132. Refer to the asynchronous strategy (15.2) and assume only that the step-size parameter $\boldsymbol{\mu}_k(i)$ follows a binomial distribution. Specifically, at every iteration i, the value assumed by $\boldsymbol{\mu}_k(i)$ is either zero with probability p or μ with probability $1-p$. Assume the neighborhoods do not change with time, as well as the combination weights (which continue to be entries from a left-stochastic combination policy A).

(a) Derive an expression for the network MSD for sufficiently small step-sizes.
(b) Derive an expression for the convergence rate of the network.
(c) How do these results compare with the performance of a synchronous network where each agent k employs $\mathbb{E}\,\boldsymbol{\mu}_k(i)$ as its step-size.