1 Essentials

Inner Product: $\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{x}^\top\mathbf{y} = \sum_{i=1}^N \mathbf{x}_i\mathbf{y}_i$ • $\langle \mathbf{x},\mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x},\mathbf{y}\rangle + \langle \mathbf{x},\mathbf{z}\rangle = \langle \mathbf{x}+\mathbf{y},\mathbf{z}\rangle$ • $\langle \alpha\mathbf{u},\mathbf{a}\rangle = \alpha\langle\mathbf{u},\mathbf{a}\rangle$

orthogonal: $\mathbf{A}^{-1} = \mathbf{A}^\mathbf{T}$ • $det(\mathbf{A}) \in +1/-1$ • $det(\mathbf{A}\mathbf{A}^\mathbf{T}) = 1$ • $\mathbf{U_K}\mathbf{U_K^T}\mathbf{U} = [\mathbf{U_K};\mathbf{0}]$ • $\mathbf{U}: orth. \to \mathbf{U}^\top orth.$ • $\|\mathbf{Ux}\|_2 = \|\mathbf{x}\|_2$ • $\|\mathbf{A}\|_2 = \max_\mathbf{x}\|\mathbf{Ax}\|_2 = \max_\mathbf{x}\|\mathbf{U}(\mathbf{DV}^\top\mathbf{x})\|_2 = \max_\mathbf{x}\|\mathbf{DV}^\top\mathbf{x}\|_2 = \max_\mathbf{y}\|\mathbf{Dy}\|_2 = \|\mathbf{D}\|_2$ with $\|\mathbf{x}\|_2 = 1$ and $\mathbf{y} = \mathbf{V}^\top\mathbf{x}$ • $\langle\mathbf{u},\mathbf{v}\rangle = 0$

trace: $\text{tr}(\mathbf{XYZ}) = \text{tr}(\mathbf{ZXY})$ • $\text{tr}(c\mathbf{A}) = c*\text{tr}(\mathbf{A})$ • $\text{tr}(\mathbf{A}+\mathbf{B}) = \text{tr}(\mathbf{A})+\text{tr}(\mathbf{B})$ • $\text{tr}(\mathbf{u_i}\mathbf{u_i}^\top) = 1$

CoordDesc: $f(\mathbf{x}) = \frac12\|\mathbf{y}-\mathbf{Ax}\|^2$: $\nabla_i f(\mathbf{x}) = \frac12\frac{\partial}{\partial \mathbf{x}_i}\|\mathbf{y}-\mathbf{Ax}\|^2 = -2\frac12[\frac{\partial}{\partial\mathbf{x_i}}\mathbf{Ax}]^\top(\mathbf{y}-\mathbf{Ax}) \overset{1}{=} \mathbf{A_i}^\top(\mathbf{Ax}-\mathbf{y}) \overset{2}{=} \mathbf{A_i}^\top(\mathbf{A_ix_i}+\sum_{j\neq i}\mathbf{A_jx_j}-\mathbf{y}) \overset{!}{=} 0 \to \mathbf{x_i} = \frac{\mathbf{A_i}^\top(\mathbf{y}-\sum_{j\neq i}\mathbf{A_jx_j})}{\mathbf{A_i}^\top\mathbf{A_i}}$ 1. $\frac{\partial}{\partial\mathbf{x_i}}\mathbf{Ax} = \frac{\partial}{\partial \mathbf{x_i}}\sum_i \mathbf{A_ix_i} = \mathbf{A_i}$ 2. $\mathbf{Ax} = \sum_j\mathbf{A_jx_j} = \mathbf{A_ix_i}\sum_{j\neq i}\mathbf{A_jx_j}$

EVD: $\mathbf{Y} = \mathbf{X} + \mathbf{uu}^\top$ • $\mathbf{X}$:symm • distinct λ_1 & λ_2 • EV $\mathbf{u}$ & $\mathbf{v}$ • $\mathbf{Y}$ symm: $\mathbf{Y}^\top = \mathbf{Y}$ • $\mathbf{u}$ EV of $\mathbf{Y}$: $\mathbf{Yu} = (\mathbf{X}+\mathbf{uu}^\top)\mathbf{u} = \mathbf{Xu}+\mathbf{uu}^\top\mathbf{u} \overset{\mathbf{uu}^\top=1}{=} \mathbf{Xu}+\mathbf{u} \overset{\mathbf{Xu}=\lambda_1\mathbf{u}}{=} \lambda_1\mathbf{u}+\mathbf{u} = (\lambda_1+1)\mathbf{u}$ • $\mathbf{v}$ EV of $\mathbf{Y}$: $\mathbf{Yv} = (\mathbf{X}+\mathbf{uu}^\top)\mathbf{v} = \mathbf{Xv}+\mathbf{uu}^\top\mathbf{v} \overset{\mathbf{u}^\top\mathbf{v}=0}{=} \mathbf{Xv} = \lambda_2\mathbf{v}$

1.1 Norms

• $\|\mathbf{x}\|_0 := |\{i|x_i\neq 0\}|$ • $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^N \mathbf{x}_i^2} = \sqrt{\mathbf{x}^\top\mathbf{x}}$ • $\|\mathbf{x}\|_p := (\sum_{i=1}^N|x_i|^p)^{\frac1p}$ • $\|\mathbf{X}\|_\star = \sum_{i=1}^{\min(m,n)}\sigma_i$ • $\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^M\sum_{j=1}^N|\mathbf{A}_{i,j}|^2} = \sqrt{\text{trace}(\mathbf{A}^\top\mathbf{A})} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2}$ • $\|\mathbf{X}\|_1 = \max_{1\le j\le n}\sum_{i=1}^m|x_{ij}|$ • $\|\mathbf{X}\|_2 = \sigma_{max}(\mathbf{X})$

1.2 Derivatives

$\partial f/\partial\mathbf{X}$: • $\partial\mathbf{A} = 0$ • $\partial(\alpha\mathbf{X}) = \alpha\partial\mathbf{X}$ • $\partial(\mathbf{X}+\mathbf{Y}) = \partial(\mathbf{X})+\partial(\mathbf{Y})$ • $\partial(trace(\mathbf{X})) = trace(\partial\mathbf{X})$ • $\partial(\mathbf{XY}) = \partial(\mathbf{X})\mathbf{Y}+\mathbf{X}\partial(\mathbf{Y})$ • $\partial(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1}$ • $\partial\mathbf{X}^\mathbf{T} = (\partial\mathbf{X})^T$

Vectors: • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{Ax}) = \mathbf{A}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{Ax}) = (\mathbf{A}^\top+\mathbf{A})\mathbf{x} \overset{\text{if }\mathbf{A}\text{ sym.}}{=} 2\mathbf{Ax}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{Ax}) = \mathbf{A}^\top\mathbf{b}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{s}^\top\mathbf{Ar}) = (\frac{\partial\mathbf{s}}{\partial\mathbf{x}})^\top\mathbf{Ar}+(\frac{\partial\mathbf{r}}{\partial\mathbf{x}})^\top\mathbf{A}^\top\mathbf{s}$

scalar α: • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{y}^\top\mathbf{Ax}) = \mathbf{y}^\top\mathbf{A}$ • $\frac{\partial}{\partial\mathbf{y}}(\mathbf{y}^\top\mathbf{Ax}) = \mathbf{x}^\top\mathbf{A}^\top$

Matrices: • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{b}^\top\mathbf{X}^\top\mathbf{Xc}) = \mathbf{X}(\mathbf{bc}^\top+\mathbf{cb}^\top)$ • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{Xb}) = \mathbf{cb}^\top$ • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b}) = \mathbf{bc}^\top$ • $\frac{\partial}{\partial\mathbf{U}}(\mathbf{U}^\top\mathbf{V}) = \frac{\partial}{\partial\mathbf{U}}(\mathbf{V}^\top\mathbf{U})^\top = (\frac{\partial}{\partial\mathbf{U}}\mathbf{V}^\top\mathbf{U})^\top = (\mathbf{V}^\top)^\top = \mathbf{V}$ • $\frac{\partial f}{\partial\mathbf{X}^\mathbf{T}} = (\frac{\partial f}{\partial\mathbf{X}})^T$

Norms: • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}\|_1) = \mathbf{1}$ • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}-\mathbf{b}\|_2) = \frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$ • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}^\top\mathbf{x}\|_2) = 2\mathbf{x}$ • $\frac{\partial}{\partial\mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$ • $\frac{\partial}{\partial\mathbf{W}}(\|\mathbf{XW}-\mathbf{Y}\|_F^2) = 2\mathbf{X}^\top(\mathbf{XW}-\mathbf{Y})$

1.3 Eigendecomposition

• $\mathbf{A}\in\mathbb{R}^{N\times N}$: $\mathbf{A} = \mathbf{Q\Lambda Q}^{-1}$ & $\mathbf{Q}\in\mathbb{R}^{N\times N}$ • if all $\lambda_i\neq 0$: $\mathbf{A}^{-1} = \mathbf{Q\Lambda}^{-1}\mathbf{Q}^{-1}$ and $(\mathbf{\Lambda}^{-1})_{i,i} = \frac{1}{\lambda_i}$ • if $\mathbf{A}$ symm.: $A = \mathbf{Q\Lambda Q}^\top$

1.4 Probability / Statistics

• $P(x) := Pr[X=x] := \sum_{y\in Y}P(x,y)$ • $P(x|y) := Pr[X=x|Y=y] := \frac{P(x,y)}{P(y)}$, if $P(y)>0$ • $\forall fixed y\in Y: \sum_{x\in X}P(x|y) = 1$ • $P(x,y) = P(x|y)P(y)$ • $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$ • $P(x|y) = P(x) \Leftrightarrow P(y|x) = P(y)$ (iff X, Y ind.) • $P(x_1,\ldots,x_n) \overset{IID}{=} \prod_{i=1}^n P(x_i)$

2 Dimensionality Reduction / PCA

$\mathbf{X}\in\mathbb{R}^{D\times N}$. N observations, K properties. Target: $\tilde{\mathbf{X}}\in\mathbb{R}^{K\times N}$. 1. Mean: $\bar{\mathbf{x}} = \frac1N\sum_{n=1}^N\mathbf{x}_n$ 2. Center: $\overline{\mathbf{X}} = \mathbf{X}-[\bar{\mathbf{x}},\ldots,\bar{\mathbf{x}}] = \mathbf{X}-\mathbf{M}$ 3. Cov. Matrix: $\mathbf{\Sigma} = \frac1N\sum_{n=1}^N(\mathbf{x}_n-\bar{\mathbf{x}})(\mathbf{x}_n-\bar{\mathbf{x}})^\top = \frac1N\overline{\mathbf{X}}\overline{\mathbf{X}}^\top$ 4. EVD: $\mathbf{\Sigma} = \mathbf{U\Lambda U}^\top$ 5. Select $K<D$, $\Rightarrow \mathbf{U}_K,\boldsymbol{\lambda}_K$ 6. Transform to new Basis: $\overline{\mathbf{Z}}_K = \mathbf{U}_K^\top\overline{\mathbf{X}}$ 7. Reconstruct original Basis: $\tilde{\overline{\mathbf{X}}} = \mathbf{U}_K\overline{\mathbf{Z}}_K$ 8. Reverse centering: $\tilde{\mathbf{X}} = \tilde{\overline{\mathbf{X}}}+\mathbf{M}$

• $\mathbf{U}_k\in\mathbb{R}^{D\times K},\mathbf{\Sigma}\in\mathbb{R}^{D\times D},\overline{\mathbf{Z}}_K\in\mathbb{R}^{K\times N},\overline{\mathbf{X}}\in\mathbb{R}^{D\times N}$

• error $J = \sum_{d=K+1}^D\mathbf{u}_d^\top\mathbf{\Sigma}\mathbf{u}_d$

3 SVD

• $\mathbf{A} = \mathbf{UDV}^\top = \sum_{k=1}^{\text{rank}(\mathbf{A})}d_{k,k}u_k(v_k)^\top$ • $\mathbf{A}\in\mathbb{R}^{N\times P},\mathbf{U}\in\mathbb{R}^{N\times N},\mathbf{D}\in\mathbb{R}^{N\times P},\mathbf{V}\in\mathbb{R}^{P\times P}$ • $\mathbf{U}^\top\mathbf{U} = I = \mathbf{V}^\top\mathbf{V}$ • $\mathbf{U}$ columns are EVs of $\mathbf{AA}^\top$, $\mathbf{V}$ columns are EVs of $\mathbf{A}^\top\mathbf{A}$, $\sigma = \sqrt{\lambda}$. If $\mathbf{S} = \mathbf{S}^\top \Rightarrow \mathbf{S} = \mathbf{UDU}^\top$ • Missing columns in $\mathbf{U}$ are basis of null$(A^\top)$ and in $\mathbf{V}$ are basis of null(A). • $\mathbf{U}$: users-to-concept, $\mathbf{V}$: Movies-to-concept, $\mathbf{D}$: expressiveness of concept

3.1 Low-Rank approximation

$\tilde{\mathbf{A}}_{i,j} = \sum_{k=1}^K\mathbf{U}_{i,k}\mathbf{D}_{k,k}\mathbf{V}_{j,k} = \mathbf{U}_{i,k}\mathbf{D}_{k,k}(\mathbf{V}^\top)_{k,j}$.

Error Frobenius: $\|\mathbf{A}-\tilde{\mathbf{A}}\|_F = \sqrt{\sum_{i>K}\sigma_i^2} = \sqrt{\sum_{i>K}\lambda_i}$

Error Euclidean: $\|\mathbf{A}-\tilde{\mathbf{A}}\|_2 = \sigma_{K+1}$

4 K-means Algorithm

Target: $\min_{\mathbf{U},\mathbf{Z}}J(\mathbf{U},\mathbf{Z}) = \|\mathbf{X}-\mathbf{UZ}\|_F^2 = \sum_{n=1}^N\sum_{k=1}^K z_{k,n}\|\mathbf{x}_n-\mathbf{u}_k\|_2^2$ 1. choose K centroids $\mathbf{U} = [\mathbf{u}_1,\ldots,\mathbf{u}_K]$ 2. Assign data points to clusters. $k^\star(\mathbf{x}_n) = \arg\min_k\{\|\mathbf{x}_n-\mathbf{u}_k\|_2\}$. Set $\mathbf{z}_{k^\star,n} = 1$, and for $l\neq k^\star$ $\mathbf{z}_{l,n} = 0$. 3. Update centroids: $\mathbf{u}_k = \frac{\sum_{n=1}^N z_{k,n}\mathbf{x}_n}{\sum_{n=1}^N z_{k,n}}$. 4. goto step 2, stop if $\|\mathbf{Z}-\mathbf{Z}^{\text{new}}\|_0 = \|\mathbf{Z}-\mathbf{Z}^{\text{new}}\|_F^2 = 0$.

5 Gaussian Mixture Models (GMM)

For GMM let $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)$; $p_{\theta_k}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\Sigma_k)$

Mixture Models: $p_\theta(\mathbf{x}) = \sum_{k=1}^K\pi_k p_{\theta_k}(\mathbf{x})$

Assignment variable (generative model): $z_k\in\{0,1\}$, $\sum_{k=1}^K z_k = 1$, $\Pr(z_k=1) = \pi_k \Leftrightarrow p(\mathbf{z}) = \prod_{k=1}^K\pi_k^{z_k}$

Complete data distribution: $p_\theta(\mathbf{x},\mathbf{z}) = \prod_{k=1}^K(\boldsymbol{\pi}_k p_{\theta_k}(\mathbf{x}))^{z_k}$

Posterior Probabilities: $\Pr(z_k=1|\mathbf{x}) = \frac{\Pr(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{l=1}^K\Pr(z_l=1)p(\mathbf{x}|z_l=1)} = \frac{\boldsymbol{\pi}_k p_{\theta_k}(\mathbf{x})}{\sum_{l=1}^K\boldsymbol{\pi}_l p_{\theta_l}(\mathbf{x})}$

Likelihood of observed data $\mathbf{X}$: $p_\theta(\mathbf{X}) = \prod_{n=1}^N p_\theta(\mathbf{x}_n) = \prod_{n=1}^N(\sum_{k=1}^K\pi_k p_{\theta_k}(\mathbf{x}_n))$

Maximize log-likelihood: $L(\boldsymbol{X},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = ln\,p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^N ln\{\sum_{k=1}^K\pi_k\cdot\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\}$

MLE: $\arg\max_\theta\sum_{n=1}^N\log(\sum_{k=1}^K\pi_k p_{\theta_k}(\mathbf{x}_n))$

$\log\left(\sum_{k=1}^K\frac{q_k\pi_k p_{\theta_k}(\mathbf{x}_n)}{q_k}\right) \geq \sum_{k=1}^K q_k[\log p_{\theta_k}(\mathbf{x}_n)+\log\pi_k-\log q_k]$ with $\sum_{k=1}^K q_k = 1$ by Jensen. Lagrangian and get q_k as below.

5.1 Expectation-Maximization (EM) for GMM

1. Initialize $\boldsymbol{\pi}_k^{(0)},\boldsymbol{\mu}_k^{(0)},\boldsymbol{\Sigma}_k^{(0)}$ for $k = 1,\ldots,K$ and $t = 1$.
2. E-Step: $\Pr[z_{k,n}=1|\mathbf{x}_n] = q_{k,n} = \frac{\boldsymbol{\pi}_k^{(t-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k^{(t-1)},\boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^K\boldsymbol{\pi}_j^{(t-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j^{(t-1)},\boldsymbol{\Sigma}_j^{(t-1)})}$
3. M-Step: $\boldsymbol{\mu}_k^{(t)} := \frac{\sum_{n=1}^N q_{k,n}\mathbf{x}_n}{\sum_{n=1}^N q_{k,n}}$ & $\boldsymbol{\pi}_k^{(t)} := \frac1N\sum_{n=1}^N q_{k,n}$ & $\Sigma_k^{(t)} = \frac{\sum_{n=1}^N q_{k,n}(\mathbf{x}_n-\boldsymbol{\mu}_k^{(t)})(\mathbf{x}_n-\boldsymbol{\mu}_k^{(t)})^\top}{\sum_{n=1}^N q_{k,n}}$
4. Repeat from (2.) with $t = t+1$ if not $\|\log p(\mathbf{X}|\boldsymbol{\pi}^{(t)},\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})-\log p(\mathbf{X}|\boldsymbol{\pi}^{(t-1)},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})\| < \varepsilon$

5.2 Model Order Selection (AIC / BIC for GMM)

Trade-off: data fit (likelihood $p(\mathbf{X}|\theta)$) vs complexity (#free parameters $\kappa(\cdot)$). Choosing K: • $\text{AIC}(\theta|\mathbf{X}) = -\log p_\theta(\mathbf{X})+\kappa(\theta)$ • $\text{BIC}(\theta|\mathbf{X}) = -\log p_\theta(\mathbf{X})+\frac12\kappa(\theta)\log N$ • AIC vs BIC for dif. K: smaller = better. BIC penalizes complexity more.

6 Word Embeddings

Distributional Model: $p_\theta(w|w') = \Pr[w \text{ occurs close to } w']$

Log-likelihood: $L(\theta;\mathbf{w}) = \sum_{t=1}^T\sum_{\Delta\in I}\log p_\theta(w^{(t+\Delta)}|w^{(t)})$

Latent Vector Model: $w\mapsto(\mathbf{x}_w,b_w)\in\mathbb{R}^{D+1}$

$p_\theta(w|w') = \frac{\exp[\langle\mathbf{x}_w,\mathbf{x}_{w'}\rangle+b_w]}{\sum_{v\in V}\exp[\langle\mathbf{x}_v,\mathbf{x}_{w'}\rangle+b_v]}$. • vocab V, context vocab C: $\log p_\theta(w|w') = \langle y_w,x_{w'}\rangle+b_w$, word embed. y_w, context embed. $x_{w'}$ • use GloVe objective

6.1 GloVe (Weighted Square Loss)

Co-occurence Matrix: $\mathbf{N} = (n_{ij})\in\mathbb{R}^{|V|\cdot|C|} \leftrightarrow \#w_i$ in c'txt w_j

Objective: $H(\theta;\mathbf{N}) = \sum_{n_{ij}>0}f(n_{ij})(\log n_{ij}-\log\exp[\langle\mathbf{x}_i,\mathbf{y}_j\rangle+b_i+d_j])^2$ with $f(n) = \min\{1,(\frac{n}{n_{max}})^\alpha\}$, $\alpha\in(0;1]$.

SGD: 1. $\mathbf{x}_i^{new}\leftarrow\mathbf{x}_i+2\eta f(n_{ij})(\log n_{ij}-\langle\mathbf{x}_i,\mathbf{y}_j\rangle)\mathbf{y}_j$
2. $\mathbf{y}_j^{new}\leftarrow\mathbf{y}_j+2\eta f(n_{ij})(\log n_{ij}-\langle\mathbf{x}_i,\mathbf{y}_j\rangle)\mathbf{x}_i$

7 Non-Negative Matrix Factorization (NMF) /

Context Model: $p(w|d) = \sum_{z=1}^K p(w|z)p(z|d)$

Conditional independence assumption ($*$): $p(w|d) =$

$\sum_z p(w,z|d) = \sum_z p(w|d,z)p(z|d) \stackrel{*}{=} \sum_z p(w|z)p(z|d)$
Symmetric parameterization: $p(w,d) = \sum_z p(z)p(w|z)p(d|z)$

7.1 EM for pLSA:

$x_{ij} = \#$ occurences of w_j in d_i
1. Log-Likelihood: $L(\mathbf{U},\mathbf{V}) = \sum_{i,j} x_{i,j} \log p(w_j|d_i) = \sum_{(i,j)\in X} \log \sum_{z=1}^K p(w_j|z)p(z|d_i)$
2. E-Step (optimal q): $q_{zij} = \frac{p(w_j|z)p(z|d_i)}{\sum_{k=1}^K p(w_j|k)p(k|d_i)} := \frac{v_{zj}u_{zi}}{\sum_{k=1}^K u_{kj}v_{ki}}$
3. M-Steps: $u_{zi} = p(z|d_i) = \frac{\sum_j x_{ij}q_{zij}}{\sum_j x_{ij}}, v_{zj} = p(w_j|z) = \frac{\sum_i x_{ij}q_{zij}}{\sum_{i,l} x_{il}q_{zil}}$

7.2 NMF Algorithm for quadratic cost function

• $\mathbf{X} \in \mathbb{Z}_{\geq 0}^{N\times M}$ • NMF: $\mathbf{X} \approx \mathbf{U}^\top\mathbf{V}, x_{ij} = \sum_z u_{zi}v_{zj} = \langle \mathbf{u}_i, \mathbf{v}_j \rangle$
$\min_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top\mathbf{V}\|_F^2$ s.t. $\forall i,j,z\ u_{zi}, v_{zj} \geq 0$
1. init: $\mathbf{U},\mathbf{V} = rand()$ 2. repeat for *maxIters*: 3. update $\mathbf{U}$: $(\mathbf{V}\mathbf{V}^\top)\mathbf{U} = \mathbf{V}\mathbf{X}^\top$ 4. project $u_{zi} = \max\{0, u_{zi}\}$ 5. update $\mathbf{V}$: $(\mathbf{U}\mathbf{U}^\top)\mathbf{V} = \mathbf{U}\mathbf{X}$ 6. project $v_{zj} = \max\{0, v_{zj}\}$

8 Convolutional Neural Networks

sigmoid: $s(x) = \frac{1}{1+e^{-x}}$:$\nabla_x s(x) = s(x)(1-s(x))$ **Neurons**: $F_\sigma(\mathbf{x};\mathbf{w}) = \sigma(w_0 + \sum_{i=1}^M x_i w_i)$. **Output**: linear regression; $\mathbf{y} = \mathbf{W}^L\mathbf{x}^{L-1}$, binary classification; $y_1 = \mathrm{P}[Y = 1|\mathbf{x}] = \frac{1}{1+\exp[-\langle \mathbf{w}_1^L, \mathbf{x}^{L-1}\rangle]}$, multiclass; $y_k = \mathrm{P}[Y = k|\mathbf{x}] = \frac{\exp[\langle \mathbf{w}_k^L, \mathbf{x}^{L-1}\rangle]}{\sum_{m=1}^K \exp[\langle \mathbf{w}_m^L, \mathbf{x}^{L-1}\rangle]}$. **Loss function** $l(y,\hat{y})$: squared loss; $\frac{1}{2}(y-\hat{y})^2$, cross-entropy loss; $-y\log\hat{y} - (1-y)\log(1-\hat{y})$. Convolution: $F_{n,m}(\mathbf{x};\mathbf{w}) = \sigma(b + \sum_{k=-2}^2 \sum_{l=-2}^2 w_{k,l}x_{n+k,m+l})$. • l n-channel filters K^l • channel $K_c^l, c \in [1..n]$ • pixels $(K_c^l)_{i,j}, -k \leq i,j \leq k$ • input image $(I_c)_{1\leq c\leq n}$: $(I_c)_{i,j}$ • output conv layer (i',j'): $(I \star K^l)_{i',j'} = \sum_{1\leq c\leq n}\sum_{-k\leq i,j\leq k}(I_c)_{i'+i,j'+j}(K_c^l)_{ij}$ • zero padding: $(I_c)_{a,b} = 0$ outside pixel range • non-linearity $ReLU(x) = max(0,x)$ applied per pixel • conv & ReLU: $ReLU((I\star K^l))_{i',j'}$ • per channel max-pooling (3x3) without stride, image pixel (i,j): $\max_{-1\leq i',j'\leq 1}(ReLU(I\star K^l))_{i+i',j+j'}$

9 Optimization

9.1 Unconstrained min: $\min f(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^D$

9.1.1 Coordinate Descent

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$ 2. for $t = 0$ to *maxIter*: 3. sample u.a.r. $d \sim \{1,\ldots,D\}$ 4. $u^\star = \arg\min_{u\in\mathbb{R}} f(x_1^{(t)},..,x_{d-1}^{(t)},u,x_{d+1}^{(t)},..,x_D^{(t)})$ 5. $\mathbf{x}_d^{(t+1)} = u^\star$ and $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)}$ for $i \neq d$

9.1.2 Gradient Descent (or Deepest Descent)

Gradient: $\nabla f(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial x_1},\ldots,\frac{\partial f(\mathbf{x})}{\partial x_D}\right)^\top$ 1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$ 2. for $t = 0$ to *maxIter*: $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma\nabla f(\mathbf{x}^{(t)})$, usually $\gamma \approx \frac{1}{t}$

9.1.3 Stochastic Gradient Descent (SGD)

Assume **Additive Objective**; $f(x) = \frac{1}{N}\sum_{n=1}^N f_n(x)$ 1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$ 2. for $t = 0$ to *maxIter*: 3. sample u.a.r. $n \sim \{1,\ldots,N\}$ 4. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma\nabla f_n(\mathbf{x}^{(t)})$, usually stepsize $\gamma \approx \frac{1}{t}$.

9.2 Projected Gradient Descent (Constrained Opt.)

minimize $f(x), x \in Q$ (constraint). **Project** x onto Q: $P_Q(\mathbf{x}) = \arg\min_{\mathbf{y}\in Q}\|\mathbf{y} - \mathbf{x}\|$, **Projected Gradient Update**: $\mathbf{x}^{(t+1)} = P_Q[\mathbf{x}^{(t)} - \gamma\nabla f(\mathbf{x}^{(t)})]$, $\mathbf{x}^{(t+1)}$ is unique if Q convex.

9.3 Lagrangian Multipliers

Min $f(\mathbf{x})$ st $g_i(\mathbf{x}) \leq 0,\ i:1..m$ & $h_i(\mathbf{x}) = \mathbf{a}_i^\top\mathbf{x} - b_i = 0,\ i:1..p$
Lagrangian: $L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) := f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})$
Dual function: $D(\boldsymbol{\lambda},\boldsymbol{\nu}) := \inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) \in \mathbb{R}$
Dual Problem: $\max_{\boldsymbol{\lambda},\boldsymbol{\nu}} D(\boldsymbol{\lambda},\boldsymbol{\nu})$ s.t. $\boldsymbol{\lambda} \geq \mathbf{0}$. Note: $\max_{\boldsymbol{\lambda},\boldsymbol{\nu}} D(\boldsymbol{\lambda},\boldsymbol{\nu}) \leq \min_{\mathbf{x}} f(\mathbf{x})$, equality if *dom* f and f convex

9.4 Convex Optimization

$f: \mathbb{R}^D \to \mathbb{R}$ is convex, if *dom* f is a convex set, and if $\forall \mathbf{x},\mathbf{y} \in$ *dom* f, and for $0 \leq \alpha \leq 1$: $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$. local=global min, **Convergence**: $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{c}{t}$.
Subgradient $g \in \mathbb{R}^D$ of f at $\mathbf{x}$: $f(\mathbf{y}) \geq f(\mathbf{x}) + g^\top(\mathbf{y} - \mathbf{x})\ \forall\mathbf{y}$
Epigraph of $f: \mathbb{R}^D \to \mathbb{R}: \{(\mathbf{x},t)|\mathbf{x} \in \mathrm{dom} f, f(\mathbf{x}) \leq t\}$, a fct is convex iff its epigraph is a convex set. **Convex fcts** $f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x}$; $f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x} + \mathbf{b}$; $f(\mathbf{x}) = e^{\alpha\mathbf{x}}$; Norms on $\mathbb{R}^D$

9.4.1 with Equality Constraints

• $\min_{\mathbf{x}} f(\mathbf{x})$ s.t. $\mathbf{A}\mathbf{x} = \mathbf{b}$ • $\mathscr{L}(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$
• $D(\boldsymbol{\lambda}) = \inf_{\mathbf{x}}\mathscr{L}(\mathbf{x},\boldsymbol{\lambda})$ • $\max_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda})$, $\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda})$
• Recover optimal $\mathbf{x}^* \in \arg\min_{\mathbf{x}}\mathscr{L}(\mathbf{x},\boldsymbol{\lambda}^*)$

10 Sparse Coding: $\min_{\mathbf{z}}\|\mathbf{z}\|_0$ *s.t.* $\|\mathbf{U}\mathbf{z} - \mathbf{x}\|_2 < \sigma$

Energy preserving: $\|\mathbf{U}^\top\mathbf{x}\|^2 = \|\mathbf{x}\|^2$

10.1 Orthogonal Basis

For $\mathbf{x}$ and o.n.b. $\mathbf{U}$ compute $\mathbf{z} = \mathbf{U}^\top\mathbf{x}$. Approx $\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}$, $\hat{z}_i = z_i$ if $|z_i| > \varepsilon$ else 0. Reconstruction Error $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{d\notin\sigma}\langle\mathbf{x},\mathbf{u}_d\rangle^2$.

10.2 Overcomplete Basis

$\mathbf{U} \in \mathbb{R}^{D\times L}$ and $L > D$. Decoding involved $\to$ add constraint $\mathbf{z}^\star \in \arg\min_{\mathbf{z}}\|\mathbf{z}\|_0$ s.t. $\mathbf{x} = \mathbf{U}\mathbf{z}$. NP-hard (non-convex) $\to$ approximate with 1-norm (convex, $\mathbf{z}^* = \arg\min_{\mathbf{z}}\|\mathbf{z}\|_1$ *s.t.* $\mathbf{x} = \mathbf{U}\mathbf{z}$) or with MP.
Coherence • $m(\mathbf{U}) = \max_{i,j:i\neq j}|\mathbf{u}_i^\top\mathbf{u}_j|$ • $m(\mathbf{B}) = 0$ if $\mathbf{B}$ orthogonal basis • $m([\mathbf{B},\mathbf{u}]) \geq \frac{1}{\sqrt{D}}$ if atom $\mathbf{u}$ is added to $\mathbf{B}$
Noisy observations: $\mathbf{x} = \mathbf{U}\mathbf{z} + \mathbf{n} \sim \mathscr{N}(\mathbf{0},\sigma^2\mathbf{I})$
Matching Pursuit (MP) approximation of $\mathbf{x}$ onto $\mathbf{U}$, using K entries. Objective: $\mathbf{z}^\star \in \arg\min_{\mathbf{z}}\|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2$, s.t. $\|\mathbf{z}\|_0 \leq K$ 1. init: $z \leftarrow 0, r \leftarrow x$ 2. while $\|\mathbf{z}\|_0 < K$ do 3. select atom with smallest angle $i^\star = \arg\max_i|\langle\mathbf{u}_i,\mathbf{r}\rangle|$ 4. update coefficients: $z_{i^\star} \leftarrow z_{i^\star} + \langle\mathbf{u}_{i^\star},\mathbf{r}\rangle$ 5. update residual: $\mathbf{r} \leftarrow \mathbf{r} - \langle\mathbf{u}_{i^\star},\mathbf{r}\rangle\mathbf{u}_{i^\star}$.
Recovery of MP: Coherence $m(\mathbf{U}) = \max_{i\neq j}|\langle\mathbf{u}_i,\mathbf{u}_j\rangle|$, exact recovery when: $K < \frac{1}{2}\left(1 + \frac{1}{m(\mathbf{U})}\right)$
Compressive Sensing • $\mathbf{x} \in \mathbb{R}^D$, K-sparse in o.n.b. $\mathbf{U}$. $\mathbf{y} \in \mathbb{R}^M$ with $y_i = \langle\mathbf{w}_i,\mathbf{x}\rangle$: M lin. combinations of signal; $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} = \boldsymbol{\theta}\mathbf{z}$, $\boldsymbol{\theta} \in \mathbb{R}^{M\times D}$ • Reconstruct $\mathbf{x} \in \mathbb{R}^D$ from $\mathbf{y}$; find $\mathbf{z}^\star \in \arg\min_{\mathbf{z}}\|\mathbf{z}\|_0$, s.t. $\mathbf{y} = \boldsymbol{\theta}\mathbf{z}$ (e.g. with MP). Given $\mathbf{z}$, reconstruct $\mathbf{x}$ via $\mathbf{x} = \mathbf{U}\mathbf{z}$

10.3 Dictionary Learning

Adapt the dictionary to signal characteristics. Objective: $(\mathbf{U}^\star,\mathbf{Z}^\star) \in \arg\min_{\mathbf{U},\mathbf{Z}}\|\mathbf{X} - \mathbf{U}\cdot\mathbf{Z}\|_F^2$.
Matrix Factorization by Iter Greedy Minimization 1. Coding step(column seperable): $\mathbf{Z}^{t+1} \in \arg\min_{\mathbf{Z}}\|\mathbf{X} - \mathbf{U}^t\mathbf{Z}\|_F^2$ subject to $\mathbf{Z}$ being sparse 2. Dictionary update step: $\mathbf{U}^{t+1} \in \arg\min_{\mathbf{U}}\|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2$, subject to $\forall l \in [L]: \|\mathbf{u}_l\|_2 = 1$

11 Robust PCA

Idea: Approx. $\mathbf{X}$ with $\mathbf{L}_0 + \mathbf{S}_0$, $\mathbf{L}_0$ is low-rank, $\mathbf{S}_0$ is sparse.
• $\min_{\mathbf{L},\mathbf{S}}\mathrm{rank}(\mathbf{L}) + \mu\|\mathbf{S}\|_0$, s. t. $\mathbf{L} + \mathbf{S} = \mathbf{X}$. As non-convex, change to $\min_{\mathbf{L},\mathbf{S}}\|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1$ (*not* the same in general)
• Perfect reconstruction is *not* possible if $\mathbf{S}$ is low-rank, $\mathbf{L}$ is sparse, or $\mathbf{X}$ is low-rank *and* sparse. Formally coherence: $\|\mathbf{U}^\top\mathbf{e}_i\|^2 \leq \frac{\nu r}{n}$, $\|\mathbf{V}^\top\mathbf{e}_i\|^2 \leq \frac{\nu r}{n}$, $\|\mathbf{U}\mathbf{V}^\top\|_{ij}^2 \leq \frac{\nu r}{n^2}$: $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$

11.1 Dual Ascent (Gradient Method for Dual Problem)

$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \eta\nabla D(\boldsymbol{\lambda}^t)$, $\nabla D(\boldsymbol{\lambda}) = \mathbf{A}\mathbf{x}^* - \mathbf{b}$ for $\mathbf{x}^* \in \arg\min_{\mathbf{x}}\mathscr{L}(\mathbf{x},\boldsymbol{\lambda})$
Dual Decomposition: $f(x)$ separable, $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_N]$
$\min_{\mathbf{x}}[f(\mathbf{x}) := f_1(\mathbf{x}_1) + \cdots + f_N(\mathbf{x}_N)]$ s.t. $[\mathbf{A}\mathbf{x} := \sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i]$
$\Rightarrow \mathscr{L}(\mathbf{x},\boldsymbol{\lambda}) = \mathscr{L}_1(\mathbf{x}_1,\boldsymbol{\lambda}) + \cdots + \mathscr{L}_N(\mathbf{x}_1,\boldsymbol{\lambda}) - \boldsymbol{\lambda}^\top\mathbf{b}$
$\Rightarrow \mathscr{L}_i(\mathbf{x}_i,\boldsymbol{\lambda}) = f_i(\mathbf{x}_i) + \boldsymbol{\lambda}^\top\mathbf{A}_i\mathbf{x}_i$
Dual Decomposition for Dual Ascent:
$\mathbf{x}_i^{t+1} := \arg\min_{\mathbf{x}_i}\mathscr{L}_i(\mathbf{x}_i,\lambda^t)$; $\boldsymbol{\lambda}^{t+1} := \boldsymbol{\lambda}^t + \eta^t\left(\sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^{t+1} - \mathbf{b}\right)$

11.2 Alternating Direction Method of Multipliers (ADMM)

$\min_{\mathbf{x}_1,\mathbf{x}_2} f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2)$ s. t. $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$, f_1, f_2 convex • Augmented Lagrangian: $L_p(\mathbf{x}_1,\mathbf{x}_2,\boldsymbol{\nu}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \boldsymbol{\nu}^\top(\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}) + \frac{p}{2}\|\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}\|_2^2$
• ADMM: $\mathbf{x}_1^{(t+1)} := \arg\min_{\mathbf{x}_1} L_p(\mathbf{x}_1,\mathbf{x}_2^{(t)},\boldsymbol{\nu}^{(t)})$, $\mathbf{x}_2^{(t+1)} := \arg\min_{\mathbf{x}_2} L_p(\mathbf{x}_1^{(t+1)},\mathbf{x}_2,\boldsymbol{\nu}^{(t)})$, $\boldsymbol{\nu}^{(t+1)} := \boldsymbol{\nu}^{(t)} + p(\mathbf{A}_1\mathbf{x}_1^{(t+1)} + \mathbf{A}_2\mathbf{x}_2^{(t+1)} - \mathbf{b})$ • ADDM for RPCA: $f_1(\mathbf{L}) = \|\mathbf{L}\|_\star$, $f_2(\mathbf{S}) = \lambda\|\mathbf{S}\|_1$, $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$ becomes $\mathbf{L} + \mathbf{S} = \mathbf{X}$, therefore $L_p(\mathbf{L},\mathbf{S},\boldsymbol{\nu}) = \|\mathbf{L}\|_* + \nu\|\mathbf{S}\|_1 + \langle\boldsymbol{\nu},\mathrm{vec}(\mathbf{L} + \mathbf{S} - \mathbf{X})\rangle + \frac{p}{2}\|\mathbf{L} + \mathbf{S} - \mathbf{X}\|_F^2$