Vector norms

A *norm* is a function $\|\cdot\| : V \to \mathbb{R}$ quantifying the size of a vector. It must satisfy

1) Positive scalability: $\|a \cdot \mathbf{x}\| = |a| \cdot \|\mathbf{x}\|$ for $a \in \mathbb{R}$
2) Triangle inequality: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in V$.
3) Seperability: $\|\mathbf{x}\| = 0$ implies $\mathbf{x} = 0$.

- Most common are *p-norms*: $\|\mathbf{x}\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$
- Special case is *Euclidean norm*: $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^n x_i^2}$
- The "0-*norm*" is $\|\mathbf{x}\|_0 := |\{x_i \mid x_i \neq 0\}|$

Matrix norms

We can also define norms on matrices, satisfying the properties described above. $\mathbf{A} \in \mathbb{R}^{M \times N}$:

- *Frobenius*: $\|\mathbf{A}\|_F := \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{\sum_i \sigma_i^2}$
- *p-norm*: $\|\mathbf{A}\|_p := \sup\{\|\mathbf{A}\mathbf{x}\|_p / \|\mathbf{x}\|_p\}$
- *Euclidean*: $\|\mathbf{A}\|_2 := \sup\{\|\mathbf{A}\mathbf{x}\|_2 / \|\mathbf{x}\|_2\} = \sigma_{\max}$
- *Nuclear*: $\|\mathbf{A}\|_* := \sum_i \sigma_i$

Kullback-Leibler Divergence

Divergence between discrete probability distributions P and Q: $D_{\mathrm{KL}}(P\|Q) = \sum_{\omega \in \Omega} P(\omega) \log\left(\frac{P(\omega)}{Q(\omega)}\right)$. It has the following properties: $D_{\mathrm{KL}}(P\|Q) \geq 0$; $D_{\mathrm{KL}}(P\|Q) = 0 \iff P = Q$; $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$. Since it is not symmetric, it is not a metric/distance.

Principal Component Analysis (PCA)

Orthogonal linear projection of high dimensional data onto low dimensional subspace. Objectives: 1. Minimize error $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2$ of point $\mathbf{x}$ and its approximation $\tilde{\mathbf{x}}$. 2. Preserve information: maximize variance. Both objectives are shown to be formally equivalent.

Statistics of Projected Data: Mean: sample mean $\bar{\mathbf{x}}$, Covariance: $\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^\top$

Solution: Eigenvalue Decomposition: The eigenvalue decomposition of the covariance matrix $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ contains all relevant information.

For $K \leq D$ dimensional projection space: Choose K eigenvectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ with largest associated eigenvalues $\{\lambda_1, \ldots, \lambda_K\}$.

Multivariate Normal Distribution

$$\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi)^{-\frac{k}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

Singular Value Decomposition

Theorem (Eckart-Young)**.** *Let* $\mathbf{A}$ *be a matrix of rank* R*, if we wish to approximate* $\mathbf{A}$ *using a matrix of a lower rank* K *then,* $\tilde{\mathbf{A}} = \sum_{k=1}^K d_k \mathbf{u}_k \mathbf{v}_k^\top$ *is the closest matrix in the Frobenius norm. (Assumes ordering of singular values* $d_k \geq d_{k+1}$*)*

K-Means

Given a set of data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ find *meaningful partition* of the data with a unique label for each data point.

Model Selection: decide on number of labels K.

Objective Function: $\min_{\mathbf{U},\mathbf{Z}} J(\mathbf{U}, \mathbf{Z}) = \sum_{n,k} z_{kn} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2 = \|\mathbf{X} - \mathbf{U}\mathbf{Z}\|_F^2$, s.t. $\mathbf{U} \in \mathbb{R}^{D \times K}$ and $\mathbf{Z} \in \{0,1\}^{K \times N}$, where $\sum_k z_{kn} = 1 \ \forall n$. Hard to optimize jointly, so alternate between optimizing $\mathbf{U}$ and $\mathbf{Z}$ while keeping the other fixed.

Optimal Assignment: Can minimize each column of $\mathbf{Z}$ separately. Optimum is attained by mapping to the closest centroid: $z_{kn}^*(\mathbf{U}) = [k = \operatorname{argmin}_l \|\mathbf{x}_n - \mathbf{u}_l\|_2]$.

Optimal Centroids: Compute optimal choice of $\mathbf{U}$, given assignments $\mathbf{Z}$. Continuous variables: compute partial gradient for every centroid and set to zero: $\nabla_{\mathbf{u}_k} J(\mathbf{U}, \mathbf{Z}) = -2\sum_{n=1}^N z_{kn}(\mathbf{x}_n - \mathbf{u}_k) \stackrel{!}{=} 0 \implies \mathbf{u}_k^*(\mathbf{Z}) = \frac{\sum_{n=1}^N z_{kn}\mathbf{x}_n}{\sum_{n=1}^N z_{kn}}$.

Gaussian Mixture Models (GMM)

$p_\theta(\mathbf{x}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$, $\boldsymbol{\pi} \geq \mathbf{0}$, $\|\boldsymbol{\pi}\|_1 = 1$.

Complete Data Distribution: Introduces latent variables $\mathbf{z} \in \mathbb{R}^K$: $p(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^K [\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)]^{z_k}$

Posterior Assignments: Posterior probabilities for assignments $\Pr(z_k = 1 \mid \mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \mathbf{\Sigma}_l)}$

Lower Bounding the Log-Likelihood: $\ln p_\theta(\mathbf{x}) \geq \sum_{k=1}^K q_k [\ln p(\mathbf{x}, \theta_k) + \ln \pi_k - \ln q_k]$

Mixture Model: Expectation Step: $q_k = \frac{\pi_k \ p(\mathbf{x};\theta_k)}{\sum_{l=1}^K \pi_l \ p(\mathbf{x},\theta_l)} = \Pr(z_k = 1 \mid \mathbf{x})$

Mixture Model: Maximization Step: $\pi_k^* = \frac{1}{N}\sum_{n=1}^N q_{kn}$, $\boldsymbol{\mu}_k^* = \frac{\sum_{n=1}^N q_{kn}\mathbf{x}_n}{\sum_{n=1}^N q_{kn}}$ and $\mathbf{\Sigma}_k^* = \frac{\sum_{n=1}^N q_{kn}(\mathbf{x}_n - \boldsymbol{\mu}_k^*)(\mathbf{x}_n - \boldsymbol{\mu}_k^*)^\top}{\sum_{n=1}^N q_{kn}}$

AIC and BIC: Trade-off: achieve balance between data fit — measured by likelihood $p(\mathbf{X} \mid \theta)$ — and complexity. Complexity can be measured by the number of free parameters $\kappa(\cdot)$.

Akaike Information Criterion: $-\ln p_\theta(\mathbf{X}) + \kappa(\theta)$

Bayesian Information Criterion: $-\ln p_\theta(\mathbf{X}) + \frac{1}{2}\kappa(\theta)\ln N$

Non-Negative Matrix Factorization

Solve $\mathbf{X} \approx \mathbf{U}\mathbf{Z}$ for $\mathbf{U} \in \mathbb{R}_+^{M \times K}$, $\mathbf{Z} \in \mathbb{R}_+^{K \times N}$.

pLSI: Interprets x_{mn} as $\Pr(w_m, d_n)$ (words and documents). Normalizes $\mathbf{X}$ s.t. $\sum_{m,n} x_{mn} = 1$, constrains $\sum_m u_{mk} = 1$, $\sum_{k,n} x_{kn} = 1$. zs serve as hidden topics, assumes $\Pr(w \mid d) = \sum_z \Pr(w \mid z)\Pr(z \mid d)$. Word and document are independent given topic. Tries to find $\Pr(w, d) = \sum_z \Pr(w \mid z)\Pr(d \mid z)\Pr(z)$. Done via Expectation-Maximization:

E-Step: $\Pr(z \mid, d, w) = \frac{\Pr(z)\Pr(d|z)\Pr(w|z)}{\sum_{z' \in \mathcal{Z}} \Pr(z')\Pr(d|z')\Pr(w|z')}$

M-Step: $\Pr(w \mid z) \propto \sum_{d \in \mathcal{D}} f(d, w)\Pr(z \mid d, w)$, $\Pr(d \mid z) \propto \sum_{w \in \mathcal{W}} f(d, w)\Pr(z \mid d, w)$, $\Pr(z) \propto \sum_{d,w} f(d, w)\Pr(z \mid d, w)$

Quadratic NMF: Different objective (min Frobenius norm instead of max LL), update rules: $u_{dk} \leftarrow u_{dk}(\mathbf{X}\mathbf{Z}^\top)_{dk}/(\mathbf{U}\mathbf{Z}\mathbf{Z}^\top)_{dk}$, $z_{kn} \leftarrow z_{kn}(\mathbf{U}^\top\mathbf{X})_{kn}/(\mathbf{U}^\top\mathbf{U}\mathbf{Z})_{kn}$.

Optimization

Duality for Constrained Optimization:

- Constrained Problem Formulation (Standard Form): $\min f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) \leq 0$, $h_i(\mathbf{x}) = 0$
- Unconstrained: $\min f(\mathbf{x}) + \sum_{i=1}^m I_-(g_i(\mathbf{x})) + \sum_{i=1}^p I_0(h_i(\mathbf{x}))$. I_- and I_0 are "brickwall" indicator functions.

Dual Problem:

- $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})$
- Lagrange *dual function*: $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) := \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$.
- Lagrange *dual problem*: $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ s.t. $\boldsymbol{\lambda} \geq \mathbf{0}$.

  Strong Duality: If the primal optimization problem is convex and under some additional conditions, the solution value of the dual problem is *equal* to the solution value $f(\mathbf{x}^*)$ of the primal problem.

Convexity:

- Convex Set: A set $\mathcal{Q}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$ and any $\theta \in [0,1]$, we have $\theta\mathbf{x} + (1-\theta)\mathbf{y} \in \mathcal{Q}$.
- Convex Function: $f\colon \mathbb{R}^D \to \mathbb{R}$ is convex if $\operatorname{dom} f$ is a convex set and $f(\theta\mathbf{x}+(1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x})+(1-\theta)f(\mathbf{y})$ $\forall \mathbf{x}, \mathbf{y} \in \operatorname{dom} f, \theta \in [0,1]$.
- Convex Optimization: Convex Optimization Problems are of the form $\min f(\mathbf{x})$ s.t. $\mathbf{x} \in \mathcal{Q}$ where both f is a convex function and $\mathcal{Q}$ is a convex set.

Sparse Coding

Haar Wavelets:

- Mother wavelet: $\psi(t) = \left[t \in [0, \frac{1}{2})\right] - \left[t \in [\frac{1}{2}, 1)\right]$
- Haar function: $\psi_{n,k}(t) = 2^{n/2}\psi(2^n t - k)$, $n, k \in \mathbb{Z}$

Any continuous real function on $[0,1]$ can be approximated uniformly on $[0,1]$ by linear combinations of the constant function $\mathbf{1}, \psi(t), \psi(2t), \psi(4t), \ldots, \psi(2^n t), \ldots$ and their shifted functions.

Discrete cosine transform (DCT):

- 1D DCT: $z_k = \sum_{n=0}^{N-1} x_n \cos[\frac{\pi}{N}(n+\frac{1}{2})k]$

Compressive Sensing: Main idea: acquire the set $\mathbf{y}$ of M linear combinations of the initial signal instead of the signal itself and then reconstruct the initial signal from these measurements. $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} =: \mathbf{\Theta}\mathbf{z}$, with $\mathbf{\Theta} = \mathbf{W}\mathbf{U} \in \mathbb{R}^{M \times D}$. Surprisingly given any orthonormal basis $\mathbf{U}$ we can obtain a stable reconstruction for any K-sparse, compressible signal. Two conditions: $w_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{D})$ and $M \geq cK\log(\frac{D}{K})$, $c \in \mathbb{R}$. For $M \ll D$ ill-posed, hence $\mathbf{z}^* \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{z}\|_0$, s.t. $\mathbf{y} = \mathbf{\Theta}\mathbf{z}$. NP-hard, approximate with *Matching Pursuit* or do convex relaxation with $\|\mathbf{z}\|_1$.

Coding via orthogonal transforms: Given orig. signal $\mathbf{x}$ and orthogonal matrix $\mathbf{U}$ compute change of basis $\mathbf{z} = \mathbf{U}^\top\mathbf{x}$. Truncate "small" values, giving $\hat{\mathbf{z}}$. Compute inverse transform $\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}$.

Measure performance: error $\|\mathbf{x} - \hat{\mathbf{x}}\|$ and sparsity $\|\mathbf{z}\|_0$.

Dictionary choice: Fourier Dictionary is good for "sine like" signals, Wavelet Dictionary is good for localized signals.

Overcomplete Dictionaries

More atoms than dimensions are used ($L > D$), which can result in sparser solutions but does not have a closed form solution.

Coherence: Increasing the Overcompleteness factor $\frac{L}{D}$ can increase the sparsity of the coding, but also increases the linear dependency between atoms. Coherence is a measurement for this: $m(\mathbf{U}) = \max_{i,j:i\neq j} \left|\mathbf{u}_i^\top \mathbf{u}_j\right|$. $m(\mathbf{U}) = 0$ for an orthogonal basis $\mathbf{B}$, $m([\mathbf{B}\mathbf{u}]) \geq \frac{1}{\sqrt{D}}$ if atom $\mathbf{u}$ is added to orthogonal $\mathbf{B}$.

Signal Coding: $\mathbf{U} \in \mathbb{R}^{D \times L}$ is overcomplete, so finding $\mathbf{z}$ such that $\mathbf{x} = \mathbf{U}\mathbf{z}$ is ill-posed, more unknowns than equations. Need to add sparsity constraint: $\mathbf{z}^* \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{z}\|_0$ s.t. $\mathbf{x} = \mathbf{U}\mathbf{z}$. Problem is NP-hard, can be brute-forced for small instances, needs Matching Pursuit else.

Noisy Observations: Signal might be corrupted, $\mathbf{x} = \mathbf{U}\mathbf{z} + \mathbf{n}$ with $n_d \sim \mathcal{N}(0, \sigma^2)$. Solve either $\mathbf{z}^* \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{z}\|_0$ s.t. $\|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2^2 < D\sigma^2$ or $\mathbf{z}^* \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2$ s.t. $\|\mathbf{z}\|_0 \leq K$.

Matching Pursuit (MP) Algorithm: Greedy algorithm that starts with zero vector $\mathbf{z} = \mathbf{0}$ and residual $\mathbf{r}^0 = \mathbf{x}$. At each iteration t selects atom with maximal absolute correlation to residual $d^* \leftarrow \operatorname{argmax}_d \left|\mathbf{u}_d^\top \mathbf{r}^{(t)}\right|$ and updates vectors $z_{d^*} \leftarrow z_{d^*} + \mathbf{u}_{d^*}^\top \mathbf{r}^{(t)}$, $\mathbf{r}^{(t+1)} \leftarrow \mathbf{r}^{(t)} - (\mathbf{u}_{d^*}^\top \mathbf{r}^{(t)})\mathbf{u}_{d^*}$. Stops when $\|\mathbf{z}\|_0 = K$. MP is an approximation, but recovers exact coding when $K < \frac{1}{2}(1 + \frac{1}{m(\mathbf{U})})$.

Sparse Coding for Inpainting: Define diagonal masking matrix $\mathbf{M}$, $m_{d,d} = [\text{pixel } d \text{ is known}]$, sparse coding of known parts in overcomplete dictionary $\mathbf{U}$: $\mathbf{z}^* \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{z}\|_0$ s.t. $\|\mathbf{M}(\mathbf{x} - \mathbf{U}\mathbf{z})\|_2 < \sigma$. Image reconstruction using mask: $\hat{\mathbf{x}} = \mathbf{M}\mathbf{x} + (\mathbf{I} - \mathbf{M})\mathbf{U}\mathbf{z}^*$.

Dictionary Learning

When learning the dictionary we adapt a dictionary to signal characteristics in the data, for which we have to solve a matrix factorization problem $\mathbf{X} = \mathbf{U}\mathbf{Z}$ with sparsity constraint on $\mathbf{Z}$ and atom norm constraint on $\mathbf{U}$. $(\mathbf{U}^*, \mathbf{Z}^*) \in \operatorname{argmin}_{\mathbf{U},\mathbf{Z}} \|\mathbf{X} - \mathbf{U}\mathbf{Z}\|_F^2$, objective not jointly convex over $\mathbf{U}$ and $\mathbf{Z}$ but convex in either of them when the other one is fixed.

Iterative greedy minimization:

1) Coding step: $\mathbf{Z}^{(t+1)} \in \operatorname{argmin}_{\mathbf{Z}} \left\|\mathbf{X} - \mathbf{U}^{(t)}\mathbf{Z}\right\|_F^2$, subject to $\mathbf{Z}$ being sparse and $\mathbf{U}$ being fixed.
2) Dict. update: $\mathbf{U}^{(t+1)} \in \operatorname{argmin}_{\mathbf{U}} \left\|\mathbf{X} - \mathbf{U}\mathbf{Z}^{(t+1)}\right\|_F^2$, subject to $\|\mathbf{u}_l\|_2 = 1$ for all l and $\mathbf{Z}$ being fixed.

Coding step can be done column-wise via *Matching Pursuit* and dictionary update via *K-SVD* algorithm involving a power iteration to approximate SVD solution. Dictionary learning can also be used for Speech Enhancement by learning the Speech and Noise dictionaries and then setting the noise coefficients to zero.

Robust PCA

Goal: Find a low rank representation of a matrix $\mathbf{X}$, which is corrupted by a sparse perturbation or sparse structured noise. Additive decomposition problem: $\min_{\mathbf{L},\mathbf{S}} \operatorname{rank}(\mathbf{L}) + \lambda \|\mathbf{S}\|_0$ s.t. $\mathbf{L} + \mathbf{S} = \mathbf{X}$. This problem is non-convex and thus hard to solve, convex relaxation: $\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1$ s.t. $\mathbf{L} + \mathbf{S} = \mathbf{X}$. This is *not* the same problem, but achieves the same solution under broad conditions.

Alternating Direction Method of Multiplies (ADMM): $\min_{\mathbf{x}_1,\mathbf{x}_2} f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2)$ s.t. $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$ with f_1, f_2 convex. *Augmented Lagrangian*: $L_\rho(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\nu}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \boldsymbol{\nu}^\top(\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}) + \frac{\rho}{2}\|\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}\|_2^2$, punishes violations of the constraints even more. Update steps: $\mathbf{x}_1^{(t+1)} := \operatorname{argmin}_{\mathbf{x}_1} L_\rho(\mathbf{x}_1, \mathbf{x}_2^{(t)}, \boldsymbol{\nu}^{(t)})$, $\mathbf{x}_2^{(t+1)} := \operatorname{argmin}_{\mathbf{x}_2} L_\rho(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2, \boldsymbol{\nu}^{(t)})$, $\boldsymbol{\nu}^{(t+1)} := \boldsymbol{\nu}^{(t)} + \rho(\mathbf{A}_1\mathbf{x}_1^{(t+1)} + \mathbf{A}_2\mathbf{x}_2^{(t+1)} - \mathbf{b})$.

ADMM for RPCA: Here $f_1(\mathbf{x}_1) = \|\mathbf{L}\|_*$ and $f_2(\mathbf{x}_2) = \lambda\|\mathbf{S}\|_1$, hence $L_\rho(\mathbf{L}, \mathbf{S}, \mathbf{N}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \langle \mathbf{N}, (\mathbf{L} + \mathbf{S} - \mathbf{X})\rangle + \frac{\rho}{2}\|\mathbf{L} + \mathbf{S} - \mathbf{X}\|_F^2$. Update sequence: $\mathbf{L}^{(t+1)} := \operatorname{argmin}_{\mathbf{L}} L_\rho(\mathbf{L}, \mathbf{S}^{(t)}, \mathbf{N}^{(t)})$, $\mathbf{S}^{(t+1)} := \operatorname{argmin}_{\mathbf{S}} L_\rho(\mathbf{L}^{(t+1)}, \mathbf{S}, \mathbf{N}^{(t)})$, $\mathbf{N}^{(t+1)} := \rho(\mathbf{L}^{(t+1)} + \mathbf{S}^{(t+1)} - \mathbf{X})$. Solving explicitly: $\operatorname{argmin}_{\mathbf{L}} L_\rho(\mathbf{L}, \mathbf{S}, \mathbf{N}) = \mathcal{D}_{\rho^{-1}}(\mathbf{X} - \mathbf{S} - \rho^{-1}\mathbf{N})$, $\operatorname{argmin}_{\mathbf{S}} L_\rho(\mathbf{L}, \mathbf{S}, \mathbf{N}) = \mathcal{S}_{\rho^{-1}}(\mathbf{X} - \mathbf{L} - \rho^{-1}\mathbf{N})$, where $\mathcal{S}_\tau(x) = \operatorname{sgn}(x)\max(|x| - \tau, 0)$, $\mathcal{S}_\tau(\mathbf{X})$ applies $\mathcal{S}_\tau$ to all x_{ij}, $\mathcal{D}_\tau(\mathbf{X}) = \mathbf{U}\mathcal{S}_\tau(\mathbf{\Sigma})\mathbf{V}^\top$, where $\operatorname{SVD}(\mathbf{X}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.